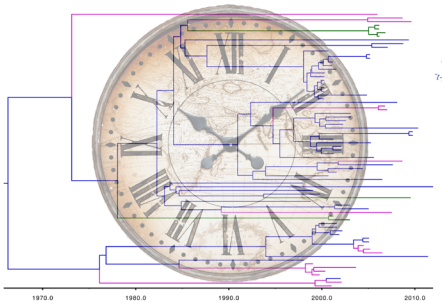
Whole Genome Epidemiological Typing of Salmonella

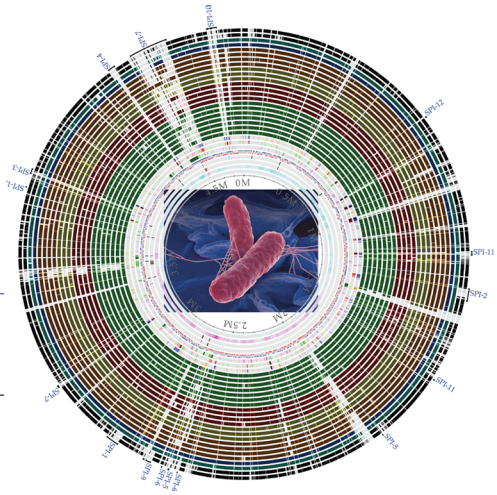
Leekitcharoenphon, Pimlapas

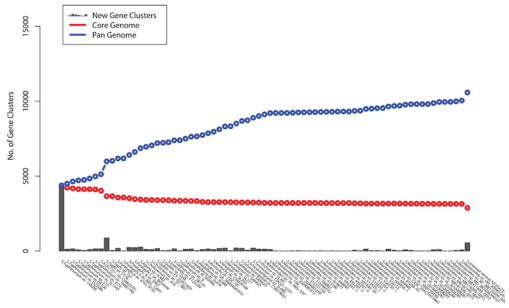
Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Whole Genome Epidemiological Typing of *Salmonella*







Pimlapas Leekitcharoenphon (Shinny)
PhD Thesis
2014

DTU Food
National Food Institute

SUPERVISORS AND FUNDING

The research has been conducted at the National Food Institute, Technical University of Denmark and Center for Biological Sequence Analysis (CBS), Technical University of Denmark. The work was supported by the Center for Genomic Epidemiology (09- 067103/DSF), http://www.genomicepidemiology.org.

Supervisors:

- Professor, PhD, **Frank Møller Aarestrup**, Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Denmark.

- Professor, PhD, **Ole Lund**, Center for Biological Sequence Analysis (CBS), Department of Systems Biology, DTU Systems Biology, Technical University of Denmark, Denmark.

- Professor, PhD, **David W. Ussery**, Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA.

Assessment Committee:

- Research area coordinator for Genomics, Dr, **Marc W Allard**, Office of Regulatory Science, Center for Food Safety & Applied Nutrition, U. S. Food & Drug Administration, MD, USA.

- Head of Typing Laboratory, PhD, **Mia Torpdahl**, Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark.

- Senior Researcher, PhD, **Henrik Hasman**, Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Denmark.

Front-page designed by Pimlapas Leekitcharoenphon and Susanne Carlsson, National Food Institute, Technical University of Denmark, Denmark.

LIST OF CONTENT

ACKNOWLEDGEMENTS

Leaving my country (Thailand) for doing PhD in Denmark and departing from +36 to -6 degree including culture differences were very challenging for me. Nonetheless, those obstacles could not be overcome without the following unforgettable persons.

First of all, I would like to express my very great appreciation to Professor Frank Møller Aarestrup for his patient guidance, enthusiastic encouragement and having faith in me by giving me opportunities to work in the awesome projects. I have learned a lot from you, more than I expected during my PhD. I wish to thank Professor Ole Lund for his valuable advices, technical support and his good humor in every weekly meeting. My grateful thanks also extended to Professor David W. Ussery for his supervision especially in the beginning of my PhD and inspiring my interest in Bioinformatics through your workshop in Thailand in 2008. Moreover you have tough me how to be a good teacher/lecturer by allowing me being teaching assistant in your workshop and courses.

My special thanks go to Rene S. Hendriksen for his research advices, giving me chances being involved in many exciting projects and importantly showing me how to communicate and corporate in research projects.

I would also like to extend my thanks to my friends/colleagues, Rolf S. Kaas, Marlene Hansen, Ana Herrero-Fresno, Carsten Friis, Simon Rasmussen, Lina Cavaco, Valeria Bortolaia, Henrik Hasman, Oksana Lukjancenko, Tammi Vesth, Maria Seier-Petersen and Mette Christiansen for their moral and/or technical supports. I also have special thanks to Rolf for his excellent programming scripts and Marlene for her help in all the Danish translation in this thesis. I would like to thank the technicians in our group, Inge Marianne Hansen, Lisbeth Andersen and Christina Svendsen who help me in many experimental works particularly genomic sequencing.

I would also like to express my grateful to another recognizable person in our group; Vibeke Dybdahl Hammer for her help in all the administrative tasks since I started my PhD and her warm welcome when I first arrived to Denmark.

Last but not least, I would like to express my gratefulness to my former teachers/ lecturers in Thailand particularly Aj. Supapon Cheevadhanarak. I also want to thank my family; mom, sisters, aunts and uncles and all of my friends both in Thailand and aboard especially Giovanni Gilardi for all of your supports throughout my PhD study.

LIST OF ORIGINAL ARTICLES

The thesis is structured as a review of a proof of concept of using WGS for epidemiological typing of *Salmonella*. Three articles that are published and one manuscript for publication in peer reviewed international journal are included in this thesis. Articles are referred in the text by roman letters and marked in **bold** typeface.

I. **Leekitcharoenphon P**, Lukjancenko O, Friis C, Aarestrup FM, Ussery DW.
Genomic variation in *Salmonella enterica* core genes for epidemiological typing. BMC Genomics. 2012 Mar 12;13:88. PMID: 22409488.

II. **Leekitcharoenphon P**, Nielsen EM, Kaas RS, Lund O, Aarestrup FM**.**
Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. PLoS One. 2014 Feb 4;9(2):e87991. PMID: 24505344.

III. **Leekitcharoenphon P**, Kaas RS, Thomsen MC, Friis C, Rasmussen S, Aarestrup FM.
snpTree--a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics. 2012;13 Suppl 7:S6. PMID: 23281601.

IV. **Leekitcharoenphon P**, Hendriksen RS, Lund O, Aarestrup FM**.**
Genomic epidemiology of the global occurrence of *S.* Typhimurium DT104. Manuscript, not submitted.

SUMMARY

Salmonella is one of the most common foodborne pathogens worldwide. In the US alone, salmonellosis was estimated to cause 1.4 million cases effecting 17,000 hospitalization and almost 600 deaths each year. Particularly, *Salmonella enterica* is a common cause of minor and large food borne outbreaks. Technological advances and effective price in high throughput genome sequencing are making whole genome sequencing (WGS) available as a routine tool for bacterial typing.

Typing of *Salmonella*, especially sub-typing within the same serotype or even the same clone, the genetic variation of the target genes being used for typing is crucial for successful discrimination. The core genes or the genes that are conserved in all members of a genus or species are potentially good candidates for investigating genomic variation in phylogeny and epidemiology. A total of 2,882 core genes have been observed among 73 available *Salmonella enterica* genomes (accessed in April 2011). A consensus tree based on variation of the core genes gives better resolution than 16S rRNA and MLST that rarely provide separation between closely related strains. The performance of the pan-genome tree which is based on the presence/absence of all genes across genomes, is similar to the consensus tree but with higher branching confidence value. The core genes can be divided into two categories: a few highly variable genes and a larger set of conserved core genes, with low variance. These core genes are useful for investigating molecular evolution and remain useful as candidate genes for bacterial genome typing-even if they cannot be expected to differentiate highly clonal isolates e.g. outbreak cases of *Salmonella* [**I**].

To achieve successful 'real-time' monitoring and identification of outbreaks, rapid and reliable sub-typing is essential. A collection of thirty-four human *S*. Typhimurium strains from six different outbreaks together with background strains plus eight *S*. Enteritidis isolates from two outbreaks and five *S*. Derby isolates from a single outbreak were used to evaluate the strengths and drawbacks of different WGS approaches compared to the traditional typing, PFGE, for retrospectively outbreak typing of *Salmonella*. The resulting outcome showed that SNP analysis and nucleotide difference approach seem to be the superior methods for outbreak detection compared to other phylogenetic analytic approaches of WGS. Furthermore, WGS approaches were also superior to the more classical typing method, PFGE. Meanwhile, k-mer method constructs a tree in high speed and giving high accuracy in clade level [**II**].

SNP analysis has successfully applied in recent epidemiological studies of *Salmonella*. Currently, there are different tools and methods to identify SNPs including various cut-off values. In addition, all the tools require bioinformatics skill. In order to apply WGS in routine typing, an automatic and user-friendly tool is needed. Therefor, snpTree has been developed as a server for online-automatic SNP analysis. snpTree can identify SNPs and construct phylogenetic tree from WGS raw reads as well as from assembled genomes or contigs. The tool is freely accessible at http://cge.cbs.dtu.dk/services/snpTree/ [**III**].

Globally, *Salmonella enterica* serovar Typhimurium is the most commonly isolated serovar. *S.* Typhimurium consists of a number of subtypes that conventionally have been divided by phagetyping. During the last three decades, *S.* Typhimurium phage type DT104 emerged as the most prevalent phage type and one of the best-studied because of its rapid global dissemination. Nonetheless, the origin and transmission route of this particular phage type have not been revealed. To bridge the gaps in epidemiology of DT104, WGS and temporally structured sequence analysis within Bayesian framework have been incorporated for reconstructing temporal and spatial phylogenies, estimating rate of mutation and divergence time of global and local *S.* Typhimurium DT104 isolates sampled from 1969 to 2012 from twenty-one countries in six continents. The DT104 was estimated to initially emerge as antimicrobial-susceptible strains in ~1946 (1931-1959) and further became multidrug-resistant (MDR) DT104 in ~1974 (1966-1981) through horizontal transfer of 13-kb SGI1 MDR region into SGI1-contained susceptible strains. Changes in population size over time supported global occurrences of MDR DT104. Besides, using WGS is capable to confirm local epidemiology especially the transmission between animal herds of DT104 isolates from Denmark. Interestingly, the demographic history of Danish MDR DT104 provided an evidence for the accomplishment of an eradicating program across pig herds in Denmark during 1996 to 2000 [**IV**].

Overall, this Ph.D. thesis has assessed the usefulness of WGS epidemiological typing in *Salmonella* as well as evaluated the different WGS approaches for outbreak investigation compared to the traditional typing, PFGE. An online tool to construct phylogenetic tree based on SNPs has also been developed. Furthermore, it has revealed the application of WGS in epidemiological study of global and local occurrences of *S.* Typhimurium DT104.

RESUMÉ

Salmonella er en af de mest almindeligt forekommende pathogene bakterier I fødevarer og fødevareproduktion på verdensplan. Alene I USA er der ca.1,4 milloner tilfælde af salmonellosis om året, hvilket resulterer i 17.000 hospitalsindlæggelser og næsten 600 dødsfald. Særligt er *Salmonella* enterica (*S*. Enterica) en hyppig årsag til både små og store udbrud af fødevareforgiftning. Teknologiske fremskridt, store prisfald og stigning I antallet af prøver der kan sekventeres samtidigt, gør nu helgenom DNAsekventering (HGS) tilgængeligt som et værktøj til rutinemæssig typing af bakterier.

I *Salmonella* typning – specielt sub-typning indenfor den same serotype eller endda samme klon, er den genetiske variation i de target gener, der bruges til typningen, afgørende for, om det lader sig gøre at adskille individuelle isolater. "Core" generne – eller de gener, der er konserverede i alle medlemmer af en genus eller art, er potientelle kandidater til at undersøge genomisk variation, udlede phylogenien og studere epidemiologien. I S. enterica er der fundet 2.882 fælles core gener i 73 offentligt tilgængelige *S*. enterica genomer (april 2011). Et konsensus træ baseret på variationen i core generne giver bedre differentiering end 16S RNA og Multi Locus Sekvens Typning (MLST), som sjældent muliggør separation af tæt beslægtede isolater. Et pan-genom træ, baseret på tilstedevær/fravær af alle gener i genomerne, gav et resultat, sammenligneligt med core konsensus, dog med højere konfidens på træets forgreninger. Core generne kan med fordel inddeles i to kategorier: En mindre gruppe gener med høj variation og en større, betående af konserverede core gener med lav varians. Disse core gener er anvendelige til studier af molekylær evolution og forbliver nyttige kandidater til bakteriel genotypning, selvom de ikke forventes at differentiere klonale isolater, eksempelvis fra et *Salmonella* udbrud [**I**].

For at opnå succefuld ,"real-time" monitorering og identifikation af udbrud er det essentielt at være i stand til at udføre hurtig og pålidelig sub-typning. En samling af 34 humane *S*. Typhimurium isolater fra seks forskellige udbrud samt baggrunds isolater, plus otte *S*. enteritidis isolater fra to udbrud og fem *S*. derby isolater fra et enkelt udbrud, blev anvendt til at evaluere styrker og svagheder i forskellige HGS baserede typninger sammenlignet med konventionel retrospekt PFGE typning af *Salmonella* udbrud. Resultatet af analyserne udpeger SNP analyse og sammenligning af parvis nukleotid forskel som superiore metoder til detektion af udbrud set i forhold til andre, ligeledes HGS baserede, phylogenetiske metoder. Overordnet set, var samtlige HGS baserede metoder superiore i forhold til den mere klassiske PFGE typning. K-mer metoden

var imidlertid den hurtigste og det resulterende klassifikations træ havde høj nøjagtighed på clade niveau [**II**].

SNP analyse har med succes været anvendt i nylige epidemiologiske studier af *Salmonella*. Aktuelt set eksisterer der forskellige værktøjer og metoder til at identificere SNPs, inklusiv adskillige, varierende cut-off værdier. Samtlige værktøjer afkræver brugeren et vist niveau af bioinformatisk kunnen. For at muliggøre anvendelsen af HGS i rutinemæssig typning er det nødvendigt at udvikle et automatisk og brugervenligt dataanalyse-værktøj. Derfor har vi udviklet snpTree – en server til automatisk, online SNP analyse. snpTree kan identificere SNPs og konstruere et phylogenetisk træ fra HGS og raw reads såvel som fra samlede genomer eller contigs. Værktøjet er frit tilgængeligt på http://cge.cbs.dtu.dk/snpTree/ [**III**].

Globalt er isolater fra *Salmonela* enterica serovar Typhimurium de hyppigst forekommende. *S.* Typhimurim består af et antal subtyper, der konventionelt er blevet opdelt via phagtypning. Gennem de forrige tre årtier, er *S.* Typhimurium phagtype DT104 vundet frem som den mest udbredte phagtype og også én af de mest studerede netop på grund af dens meget hurtige, globale spredning.

Dette til trods, har man endnu ikke været i stand til at udlede denne særlige phagtypes udspring og transmissionsrute. I et forsøg på at brolægge hullerne i epidemiologien af DT104 har vi inkorporeret HGS og temporalt struktureret Bayesian baseret sekvensanalyse til at rekonstruere temporale og geografiske phylogenier og estimerer dermed mutationsrate og divergenstidspunkt af en række globale og lokale *S.* Typhimurium DT104 isolater indsamlet over årene 1969 til 2012 fra 21 lande fordelt på seks kontinenter. Den globale spredning af DT104 er estimeret til at have oprindelse i en opblomstring af en antimikrobiel sensitiv i klon i ~1946 (1931-1959) , der via horisontal genoverførsel af en 13-kb SGI1 MR region til SGI1-positive sensitive isolater, videreudviklede sig til antimikrobiel multi-resisten (MR) DT104 i ~1974 (1966-1981). Ændringer i populations størrelsen over tid understøttede ligeledes den globale forekomst af MR DT104. Udover emergens studier, er HGS også anvendeligt til bekræftelse af lokal epidemiologi – specielt transmission af DT104 mellem danske dyre besætninger. Endvidere var det meget interessant at den demografiske historie af de danske MR DT104 gav evidens for at det danske bekæmpelsesinitiativ i svinebesætningerne i årnene 1996 til 2000 har været en succes [**IV**].

Dette Ph.D. studie har vurderet anvendeligheden af HGS baseret epidemiologisk typning af *Salmonella* samt evalueret de forskellige HGS data analyse metoder med henblik på udbruds

studier i sammenligning med traditionel PFGE typning. Et online-værktøj til at konstruere fylogenetisk træ baseret på SNPs er også blevet udviklet. Endvidere er HGS anvendt i et epidemiologisk studie der der kortlægger den lokale og globale forekomst af *S*. Typhimurium DT104.

THAI ABSTRACT (บทคัดย่อ)

Salmonella เป็นหนึ่งในเชื้อก่อโรคในอาหารที่สำคัญ ในสหรัฐอเมริกา ผู้มีอาการจากเชื้อ *Salmonella* (Salmonellosis) มีสูงถึง 1.4 ล้านคน ซึ่ง 17,000 คน ต้องเข้ารับการรักษาในโรงพยาบาลและกว่า 600 คนที่เสียชีวิตต่อปี โดยเฉพาะอย่างยิ่ง *Salmonella enterica* ซึ่งก่อให้เกิดการระบาดทั้งขนาดเล็กและขนาดใหญ่ เทคโนโลยีที่ก้าวหน้าทางด้านการถอดรหัสพันธุกรรมและราคาที่ลดลงอย่างต่อเนื่อง รวมทั้งประสิทธิภาพของการถอดรหัสพันธุกรรมได้ทำให้การถอดรหัสพันธุกรรมของ DNA ทั้งหมดในแบคทีเรียสามารถทำได้และสามารถใช้เป็นเครื่องมือในการบ่งบอกลักษณะและจำแนกแบคทีเรีย (bacterial typing).

การจัดจำแนกแบคทีเรีย (bacterial typing) โดยเฉพาะอย่างยิ่งการจำแนกย่อยในระดับ serotype และระดับ clonal ความแปรผันในระดับ DNA ของ gene เพื่อใช้ในการจำแนกเป็นสิ่งที่สำคัญมาก สำหรับการจำแนกที่มีประสิทธิภาพ core genes หรือ genes ที่พบในทุกๆ genus หรือ species เป็นสิ่งหนึ่งที่ใช้สำหรับการศึกษาความแปรผันใน DNA สำหรับการจำแนกและระบาดวิทยาของแบคทีเรีย core genes จำนวน 2,882 genes ได้ถูกค้นพบในกลุ่มของ *Salmonella enterica* จำนวน 73 genomes phylogenetic tree ที่สร้างโดยใช้ความแปรผันทาง DNA ของ core genes ได้แสดงประสิทธิภาพในการจำแนก *Salmonella* ได้ดีกว่า phylogenetic tree จาก 16S rRNA และ MLST pan-genome tree ซึ่งสร้างโดยหลักการปรากฏและไม่ปรากฏของ genes ใน genomes ต่างๆ ของ *Salmonella* ได้แสดงประสิทธิภาพในการจำแนก *Salmonella* ได้เท่าเทียมกับ phylogenetic tree จาก core genes แต่ต่างกันตรงที่ให้ค่าความมั่นใจที่สูงกว่า core genes สามารถแบ่งได้เป็น 2 ประเภทคือ genes ที่มีความแปรผันทาง DNA สูง ซึ่งมีจำนวนน้อย และ genes ที่มีความแปรผันทาง DNA ต่ำ ซึ่งมีจำนวนมากและมักพบในทุกๆ genomes core genes เหล่านี้มีประโยชน์สำหรับการศึกษาวิวัฒนาการในระดับโมเลกุลและยังมีประโยชน์สำหรับใช้เป็น target genes ในการจำแนกแบคทีเรีย แม้ว่าจะไม่สามารถจำแนกแบคทีเรียได้ในระดับ clonal อย่างเช่น ในระดับ outbreak ของ *Salmonella* [**I**]

เพื่อที่จะได้การตรวจสอบและการบ่งชี้ outbreak แบบ real-time วิธีการจำแนกแบคทีเรีย หรือ sub-typing ที่รวดเร็วและถูกต้องเป็นสิ่งที่จำเป็น *S.* Typhimurium strains จำนวน 34 ตัวอย่างจากตัวอย่างผู้ป่วยที่สุ่มจาก 6 outbreaks รวมทั้ง background strains และ 8 *S.* Enteritidis จาก 2 outbreaks และ 5 *S.* Derby จาก 1 outbreak ได้ถูกนำมาใช้เพื่อเปรียบเทียบประสิทธิภาพ จุดเด่นและจุดด้อยของวิธีการทาง Whole genome sequencing และวิธีแบบดั้งเดิม เช่น PFGE เพื่อใช้ใน

การจำแนก *Salmonella* ในสถานการณ์ของ outbreak ผลการทดลองได้แสดงว่า การวิเคราะห์โดยใช้ SNP และ nucleotide difference เป็นวิธีที่มีประสิทธิภาพในการจำแนก outbreak strains ได้ดีกว่าวิธีทาง phylogenetic อื่นๆ มากกว่านั้นวิธีทาง WGS มีประสิทธิภาพสูงกว่าวิธีดั้งเดิมอย่าง PFGE และวิธีการจำแนกด้วย k-mer สามารถสร้าง phylogenetic tree ด้วยความเร็วสูงและมีความถูกต้องสูงในระดับ clade [**II**]

วิธีการวิเคราะห์ข้อมูล DNA โดยใช้ SNP ได้ถูกนำมาประยุกต์ใช้ในการศึกษาการระบาดของเชื้อ *Salmonella* อย่างประสบผลสำเร็จ ปัจจุบันมีโปรแกรมต่างๆ ที่สามารถตรวจสอบหา SNP ใน bacterial genome แต่โปรแกรมทั้งหมดต้องใช้ทักษะทาง bioinformatics เพื่อที่จะประยุกต์ใช้ WGS สำหรับ typing แบคทีเรีย โปรแกรมที่อัตโนมัติและง่ายต่อการใช้เป็นสิ่งจำเป็น ดังนั้น snpTree ได้ถูกพัฒนาขึ้นเพื่อเป็น web tool สำหรับวิเคราะห์ SNP แบบออนไลน์ snpTree สามารถตรวจหา SNP และสร้าง phylogenetic tree จากข้อมูล WGS แบบ raw reads และ assembled genomes หรือ contigs web tool snpTree เป็นโปรแกรมออนไลน์ที่ปราศจากค่าใช้จ่ายใดๆ http://cge.cbs.dtu.dk/services/snpTree/ [**III**].

Salmonella enterica serovar Typhimurium เป็น serovar ที่พบมากที่สุด *S.* Typhimurium ประกอบด้วย subtypes หลาย subtypes ซึ่งจัดจำแนกโดยวิธี phagetyping สามทศวรรษที่ผ่านมาพบว่า *S.* Typhimurium phage type DT104 ได้อุบัติขึ้นและกลายเป็น phage type หนึ่งที่พบมากที่สุด และเป็นตัวอย่าง phage type ที่ดีสำหรับการศึกษาการระบาดระดับโลก อย่างไรก็ตามต้นกำเนิดและเส้นทางการแพร่กระจายของเชื้อ DT104 ยังคงเป็นสิ่งที่ยังหาคำตอบไม่ได้ เพื่อที่จะหาคำตอบเหล่านี้ WGS และ temporally structured sequence analysis ด้วย Bayesian framework ได้ถูกนำมาใช้เพื่อสร้าง phylogenetic tree ที่รวบรวมข้อมูลทั้งเวลาและสถานที่ รวมทั้งประมาณอัตราการกลายพันธุ์และเวลาที่ diverse ของ DT104 ในระดับโลกและระดับท้องถิ่น ตัวอย่าง DT104 ได้ถูกสุ่มจากตัวอย่างตั้งแต่ปี 1969 ถึง 2012 จาก 21 ประเทศใน 6 ทวีป ผลการทดลองพบว่า DT104 ได้ถูกประมาณว่าอุบัติเริ่มแรกเป็นเชื้อไม่ดื้อยาปฏิชีวนะในปี ~1946 (1931-1959) และต่อมาได้กลายพันธุ์เป็นเชื้อดื้อยา (MDR) ในปี ~1974 (1966-1981) โดยผ่านการส่งผ่านแบบ horizontal ของ MDR region ขนาด 13-kb ไปสู่เชื้อไม่ดื้อยาที่มี SGI1 แบบแผนการเปลี่ยนแปลงประชากรเทียบกับเวลาได้สนับสนุนการแพร่กระจายของเชื้อ MDR DT104 ในระดับโลก อีกทั้งการประยุกต์ใช้ WGS ยังสามารถยืนยันการระบาดและการแพร่กระจายของเชื้อ DT104 ระหว่างฟาร์มปาศุสัตว์ในระดับท้องถิ่นในประเทศเดนมาร์ก น่าสนใจเป็นอย่างยิ่งที่แบบแผนการ

เปลี่ยนแปลงประชากรของเชื้อ MDR DT104 ในประเทศเดนมาร์ก สามารถใช้เป็นหลักฐานแสดงความสำเร็จของโปรแกรมการกำจัดเชื้อ DT104 (eradicating program) ระหว่างฟาร์มสุกรที่เริ่มต้นในปี 1996 ถึงปี 2000 ในประเทศเดนมาร์ก [**IV**]

โดยสรุป วิทยานิพนธ์ปริญญาเอกเล่มนี้ ได้แสดงให้เห็นถึงประโยชน์ของการประยุกต์ใช้ WGS สำหรับการจัดจำแนกเชื้อ *Salmonella* ในการศึกษาระบาดวิทยา อีกทั้งยังทำการเปรียบเทียบประสิทธิภาพของเทคนิคต่างๆ ของ WGS เพื่อใช้ในการจัดจำแนก outbreak strains เปรียบเทียบกับวิธีดั้งเดิมอย่าง PFGE โปรแกรมออนไลน์สำหรับสร้าง phylogenetic tree จาก SNP ได้ถูกพัฒนาขึ้น มากกว่านั้นวิทยานิพนธ์นี้ได้แสดงให้เห็นถึงประสิทธิภาพของการใช้ WGS สำหรับการศึกษาการแพร่กระจายในระดับโลกและท้องถิ่นของเชื้อ *S*. Typhimurium DT104

BACKGROUND

Salmonella is one of the most important food-borne bacterial pathogens, which effects both human health and food industries. Furthermore, they can spread worldwide across border of countries. Therefore, the emergence of *Salmonella* in one nation might cause problems in several countries.

The cost and time of whole genome sequencing have decreased dramatically. The technology has recently been applied successfully in various bacterial epidemiological studies including the study of some *Salmonella* sub-types. Promisingly, WGS is on the front line to be incorporated in clinical microbiology, routine typing and outbreak investigation. Prior to implementing WGS in epidemiological typing of *Salmonella*, the specific criteria to distinguish whether the isolates belong to the same clonal/outbreak are needed.

In Europe and North America, *Salmonella enterica* serovar Typhimurium is one of the most prevalent serovar of *Salmonella*. Particularly, *S.* Typhimurium DT104 that rapidly disseminated globally during 1990s. However, the origin and transmission routes are still unknown. Thus, further investigation and elucidation of the occurrence, international spread, and global epidemiology of *Salmonella* serovars and specific clones would suggest any potential monitor and strategies for prevention and control of similar successful clones.

PURPOSE

The purpose of the PhD project was to evaluate whole genome sequencing approaches for epidemiological typing and outbreak investigation of *Salmonella* as well as to apply WGS in spatial-temporal analysis of global and local occurrence of *S.* Typhimurium DT104.

RESEARCH APPROACH

The projects were derived from the activities of the Center for Genomic Epidemiology (CGE) (www.genomicepidemiology.org), to provide a proof of concept of using whole genome sequencing in bacterial epidemiology.

The specific studies conducted during this PhD project focused on the following objectives:

1. To evaluate of using WGS for epidemiological typing of *Salmonella.*

2. To evaluate different WGS approaches for outbreak investigation of *Salmonella enterica*.
3. To apply WGS for genomic epidemiological study of the global and local occurrence of *S.* Typhimurium DT104 as well as population structure, demographic history and evolution of this phagetype.

INTRODUCTION

Salmonella is a common cause of infectious disease in human and animals. It is one of the most common foodborne pathogens worldwide [1]. *Salmonella* is a genus of rod-shaped, gram-negative, non-spore forming, predominantly motile bacteria belonging to the family *Enterobacteriaceae* [2]. Previously, The genus *Salmonella* was considered as a single species, known as *Salmonella choleraesuis* [3]. The species *S.choleraesuis* confused with the *Salmonella* serotype Choleraesuis. Therefore, the novel name "*Salmonella enterica*" has been used as a replacement of the name "*Salmonella choleraesuis*" [3]. The *Salmonella* nomenclature of today was proposed using the analysis of somatic and flagella antigens by The Kauffman-White Scheme since 1980 [4,5]. The current nomenclature of *Salmonella* are defined and maintained by the World Health Organization (WHO) Collaborating Centre for Reference and Research on *Salmonella* at the Pasteur Institute, Paris, France (WHO Collaborating Centre) [6]. Currently, the genus *Salmonella* is classically divided into the species *S.bongori* and *S.enterica*; the latter further divided into six subspecies - *Salmonella enterica* subsp. *enterica* (I), *Salmonella enterica* subsp. *salamae* (II), *Salmonella enterica* subsp. *arizonae* (IIIa), *Salmonella enterica* subsp. *diarizonae* (IIIb), *Salmonella enterica* subsp. *houtenae* (IV), and *Salmonella enterica* subsp. *indica* (VI). The missing subspecies V was formally classified as *S.bongori* [7]. *Salmonella enterica* subsp. *enterica* (I) contains more than 2,500 different serotypes [8] which are primarily named by the geographical origin such as *S.* Amsterdam, *S.* Panama, and *S.* Montevideo whereas the serovars of the remaining five subspecies are named by antigenic formula [4,5].

The species *S. bongori* is predominantly associated with cold-blooded animals [9] whereas *S. enterica* is found in reptiles and warm-blooded vertebrates. Most subspecies in *S. enterica* are commonly found in reptiles, and are not often causing disease, but subspecies I representing far more serovars than the others, are typically isolated from mammals or birds and only a small fraction of serovars within subspecies I is pathogenic [10]. Most serovars are not pathogenic in their natural hosts however a range of serovars can cause disease in domestic animals, and some serovars are specific to a particular host [11].

Within the subspecies I, *S. enterica* serovar Typhimurium and serovar Enteritidis are host non-specific serovars and the most common pathogenic serovars, causing disease in a wide range of hosts [12,13] and they are generally associated with a relatively mild enteric disease [14]. In contrast, the host restricted serovars are the serovars that found in their host range and cause

severe disease in only one host for example *S. enterica* serovar Typhi and serovar Paratyphi are human-restricted, causing typhoid and paratyphoid fever respectively [15]. The host adapted serovars is associated predominately with disease in one species but may infect a limited number of other hosts such as the bovine-adapted *S. enterica* serovar Dublin may occasionally cause disease in other animals, including humans and sheep [16,17] and the porcine-adapted *S. enterica* serovar Choleraesuis are infrequently found in humans but causing severe disease [18–20].

Non-typhoidal *Salmonella* normally causes gastroenteritis, bacteremia, and subsequent focal infection [3] effecting an estimated 93.8 million cases of gastroenteritis globally each year, including 155,000 deaths [21]. Most human infections are self-limiting, nonetheless, approximately 5% of all patients infected with non-typhoidal *Salmonella* develop bacteremia. The very young, the elderly, the malnourished, or people with underlying diseases such as malaria or HIV are at significantly higher risk of developing bacteremia compared to other healthy individuals [22]. Severe infections with non-typhoidal *Salmonella* are relatively rare in Europe and North America but invasive non-typhoidal *Salmonella* is endemic in sub-Saharan Africa [23–26]. In contrast, typhoidal *Salmonella* for instance serovars Typhi and Paratyphi A cause enteric fever exclusively in humans [1]. Typhoid fever remains a severe disease in several regions in Asia, Africa and South America, whereas the disease is relatively rare in developed countries. The global burden was estimated to be more than 21 million cases and 200,000 deaths in 2000 [27].

EPIDEMIOLOGY OF *SALMONELLA*

The previous study showed that in all regions except the Oceania and North American, *S.* Enteritis and *S.* Typhimurium ranked as the first and second most common serovars respectively [28]. In Europe, the surveillance data between 2006 and 2007 showed that *S.* Enteritidis was ranked first, but decreasing, meanwhile *S.* Typhimurium was fairly consistent over time and ranked second [29]. Besides, *S.* Infantis was ranked third followed by *S.* Virchow, *S.* Newport and *S.* Hadar. In South America, *S.* Typhimurium was ranked first and *S.* Enteritidis ranked second in 2008. In addition, *S.* Isangi was highly frequent and ranked third followed by *S.* Dublin and *S.* Virchow [29]. The same pattern of the ranking of *S.* Enteritidis and *S.* Typhimurium was observed since 1997 in the United States [30], in China between 2006 and 2007 [31], and Taiwan between 1998 to 2002 [32]. The distribution of serovars in Southeast Asia is slightly different

from the global trend for example in the Philippines, Hong Kong, and Sri Lanka, *S*. Typhimurium was ranked before *S*. Enteritidis whereas it was the opposite in Singapore, South Korea, and Thailand [33,34]. The global distribution of *Salmonella* serovars in humans is influenced by various factors such as animal and environmental reservoirs and complex routes of transmission [3,10,11,35][**IV**].

Salmonella is also a zoonotic bacterium and has reservoirs in various animals. The most common domesticated animal hosts are chickens, pigs, and cattle. *Salmonella* can contaminate meat during slaughter and it can survive in fresh meats and meat products that are not thoroughly heated. Therefore, animal products are incorporate as a main vehicle of transmission. Egg is considered as another vehicle of transmission especially the contamination on the surface or in the interior of the egg [3]. Another source of human *Salmonella* infection is vegetables that are contaminated with animal manure during growing or processing.

After *Campylobacter*, *Salmonella* is the most commonly isolated bacterial pathogen found in the diagnosis of diarrhea [3]. Most of *Salmonella* infected cases are foodborne particularly human infections that are acquired from contaminated meats due to inadequate cooking or poor kitchen hygiene [36]. Acquisition of *Salmonella* from pets e.g. reptiles and birds, direct personal contact, nosocomial transmission, and waterborne transmissions are minor modes of transmission [3]. The increasing import of cheaper food products from countries with little or no programs of foodborne pathogens is another factor causing *Salmonella* infections [37]. Various studies also revealed that international travel to certain destinations is associated with relatively high risk of human salmonellosis [38–40].

TYPING OF *SALMONELLA*

Typing, by definition, means phenotypic and/or genetic analysis of bacterial isolates, below the species/subspecies level. The aim of typing is to generate strain/clone-specific fingerprints or datasets that can be used, for example, to detect or rule out cross-infections, elucidate bacterial transmission patterns and find reservoirs or sources of infection in humans. ‘Subtyping’ is often used as an exchangeable term for typing [41].

Bacteria replicate and preserve in ecological niches called reservoirs and the transmission of bacteria from any reservoirs cause the clusters of colonization or infection among humans. Those clusters are recognized as outbreak and may be considered as major epidemics if the outbreaks

are uncontrollable. Bacterial epidemiological typing detects isolate-specific genotypic or phenotypic characters or patterns that can be elucidate the sources and routes of dissemination of bacteria [42]. Bacterial typing is useful for studying the spread and population dynamics in clinical and environmental settings. Evaluation and validation of typing methods should be based on the following criteria; the stability of the markers assessed by the typing method (stability), the ability of the method to assign a type to all tested isolates (typeability), the ability of the method to assign a different type to two unrelated strains (discriminatory power), the concordance of the typing results to the available epidemiological information (epidemiological concordance), the ability of the typing method to assign the same type to a tested isolate on independent occasions (reproducibility), cost, and time consuming [41].

Serotyping

Salmonella serovars are classified using The Kauffman and White scheme which is based on antisera to two highly variable surface antigens called the O antigens and the H antigens that represent variation in the exposed part of the lipopolysaccharide and variation in the major protein of the flagellum respectively [5,43]. Serotyping is traditionally important method for *Salmonella* nomenclature [6].

Phage typing

Salmonella strains within a particular serovar can be divided into a number of phage types. The characterization of phage typing is carried out based on the pattern of phage lysis of wild strains. Phage typing is useful for epidemiology and surveillance for instance some phage type (DT204 and DT104) have a broad host range and are distributed worldwide, while other phage types (DT2 and DT99) are frequently associated with pigeons [44].

A phenotype may not always reflect evolutionary history because of the rate of genetic exchange within bacterial species. Isolates that are identical regarding to phage typing might in fact be quite unrelated, and isolates that show quite different phenotypes might be closely related [41]. Discrimination is therefore variable, typeability often partial, and reproducibility poor [41]. Phage typing is also labour-intensive and require skills and methodologies that are difficult to maintain at standard levels [41].

Pulsed-field gel electrophoresis (PFGE)

PFGE is an electrophoretic technique to separate large DNA molecules (10 kb – 10 Mb). Discrimination of isolates is based on the banding patters of PFGE [45]. PFGE has been a conventional typing method for *Salmonella* outbreak investigations and epidemiological studies [46–48]. Although this method has been widely used, the PFGE has limitations for example it is time and labour consuming [49].

Multiple-locus variable number tandem repeat analysis (MLVA)

Variable number tandem repeats (VNTR) are repeated DNA sequences that vary in copy number and are distributed widely in bacterial genomes [50–52]. VNTR has rapid evolution and considered as an important source of genetic polymorphism for strain typing [53–55]. The typing by MLVA is a PCR-based genotyping based on the polymorphic analysis of multiple VNTR loci on the chromosome [56,57]. MLVA profile is determined based on the number of repeats in each VNTR locus by PCR amplification [45]. The MLVA profile is applicable in potential outbreak situation and population studies of *Salmonella* [58,59].

WHOLE GENOME SEQUENCING

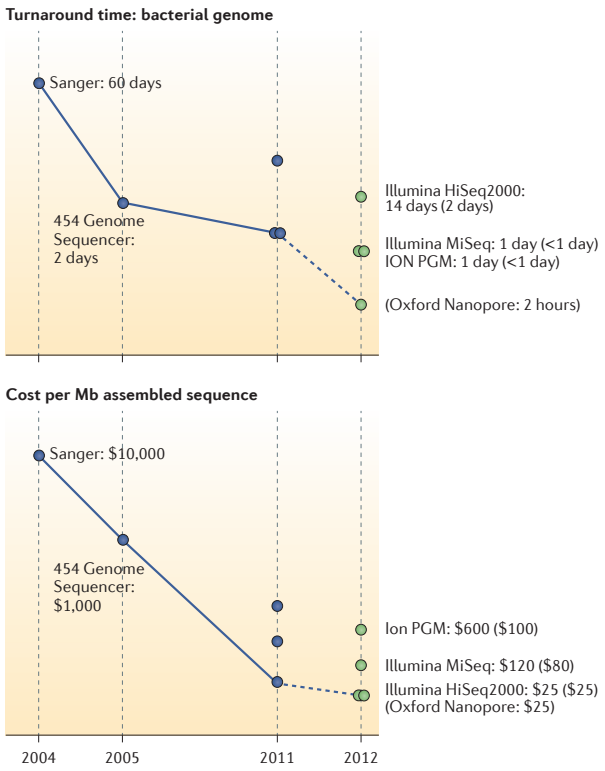


Figure 1 Comparison of sequencing performance, time consuming and cost of different WGS platforms [60].

There are various platforms or technologies to determine the complete DNA sequence of a bacterial genome (Whole genome sequencing, WGS) (Figure 1). The first next generation sequencer for WGS using pyrosequencing approaches, Roche-454, was launched in 2005 with reads of 100 bp up to 500 bp in the later versions [60,61].

Released in 2006 with 36 bp short reads, Illumina Genome Analysers are based on the Solexa sequencing-by-synthesis chemistry [62]. The lasted performance at the end of 2011 can provide ~300 Gb of raw data per eight-lane flow cell in the form of 100 bp paried-end reads. For large bacterial sample collection, the Illumina HiSeq platform is useful and cost-effective by allowing 96 samples to be sequenced simultaneously in each lane. The most popular platforms in microbiology for fast and compact bench-top machines will be the Ion PGM and the Illumina MiSeq [61].

The Ion Torrent Personal Genome Machine (PGM) was launched in early 2011 [63]. The platform incorporates a sequencing-by-synthesis using native dNTP chemistry and relying on a modified silicon chip to detect hydrogen ions released during base incorporation by DNA polymerase [61]. The new promising platform, Oxford Nonopore Technology, was planned to be released in 2012 [64] but up until now, it have not been launched. The platform allows the sequence of a single DNA molecule passing through a protein nanopore under the control of an enzyme. Nucleotide detection is measured as fluctuation in electrical current across a lipid membrane. The Oxford Nanopore sequences native DNA. It is therefore capable to apply for fairly crude sample and low DNA concentration [60]. According to the information from the company, sequence data are collected in real time at ~200 – 400 bp per second, and they expect up to 1,000 bp per second in the future and the technology has a 4% error rate [60,64].

WHOLE GENOME SEQUENCE TYPING

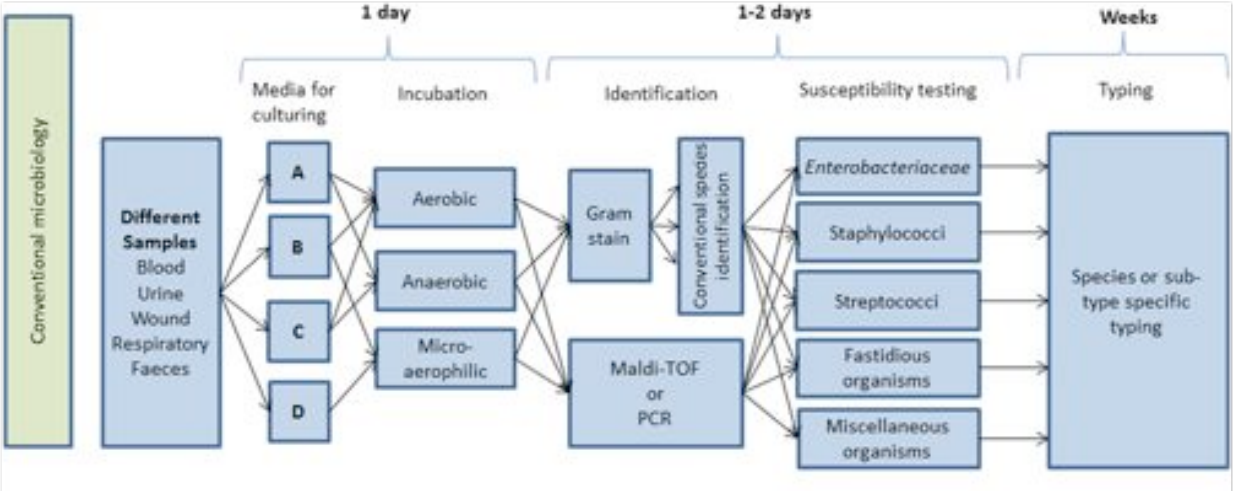


Figure 2 the principle process and timeline of traditional bacterial typing [65].

The conventional workflow of traditional typing for bacteria is illustrated in Figure 2. After culturing bacteria, the characterization of pathogens is relied on morphological appearance and density of growth that require specialist knowledge to take decision. The suspected bacterial pathogens are then processed to a complex pathway to determine species and antimicrobial susceptibility, which are based on two approaches; Gram staining and matrix-assisted laser desorption/ionization-time of light (MALDI-TOF). Eventually, a small subset of isolates may be chosen, depending on the species and ability to be part of an outbreak, for further investigation using a wide range of typing tests that are often provided by reference laboratories [60].

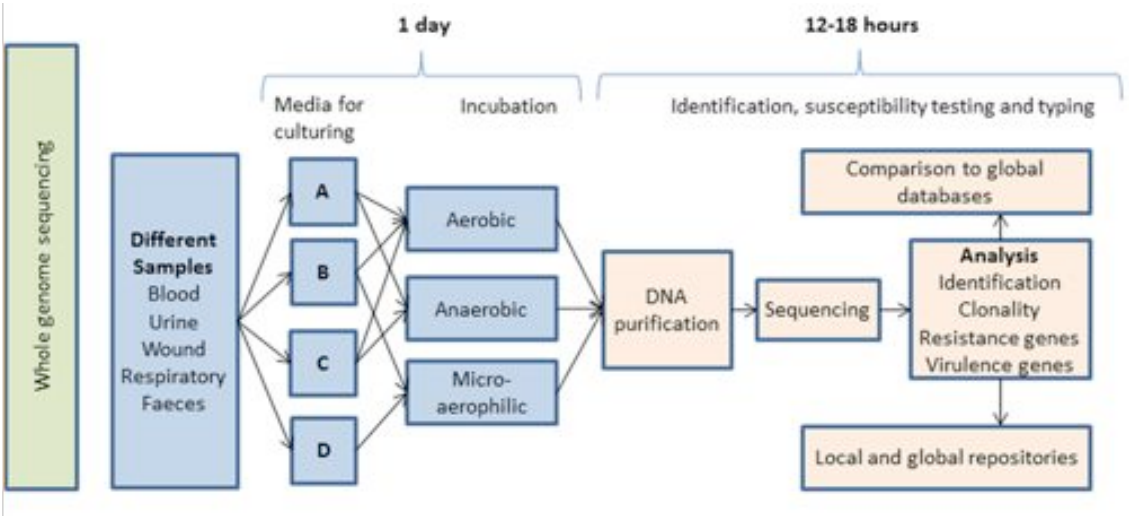


Figure 3 the principle workflow and timeline of whole genome sequence typing [65].

In contrast, whole genome sequence is a transformation of the conventional bacterial epidemiological typing [60,66]. The workflow of WGS for typing is lesser complexity and time-consuming (Figure 3). After sequencing, the main consuming processes will be computational parts. The significant advantage of WGS is to provide all of the DNA information content of isolates in a single rapid step following culturing bacteria. Fundamentally, all of the data that are currently used for diagnostic and typing can be obtained from WGS [60]. Results from WGS can be reported through an information system and will be useful for outbreak detection and infectious disease surveillance [60,65–67][**II**][**IV**]. Nonetheless, it requires a new large database, automatic tools and other informatics technologies to develop the mentioned pipeline.

COMPARATIVE GENOMIC OF *SALMONELLA*

Comparative genomics analysis of *Salmonella* genomes provides insight into the relationship between the different serovars as well as different subspecies. In principle, *Salmonella* showed fairly high similarity in protein sequences when visualized by BLAST Matrix (Figure 4), which exhibits the number of gene families found in common between the *Salmonella* genomes by pairwise all-against-all BLAST comparison at the amino acid sequences. The similarity between the proteins within *Salmonella* subsp. *enterica* ranged from 65 % to 99 % [68].

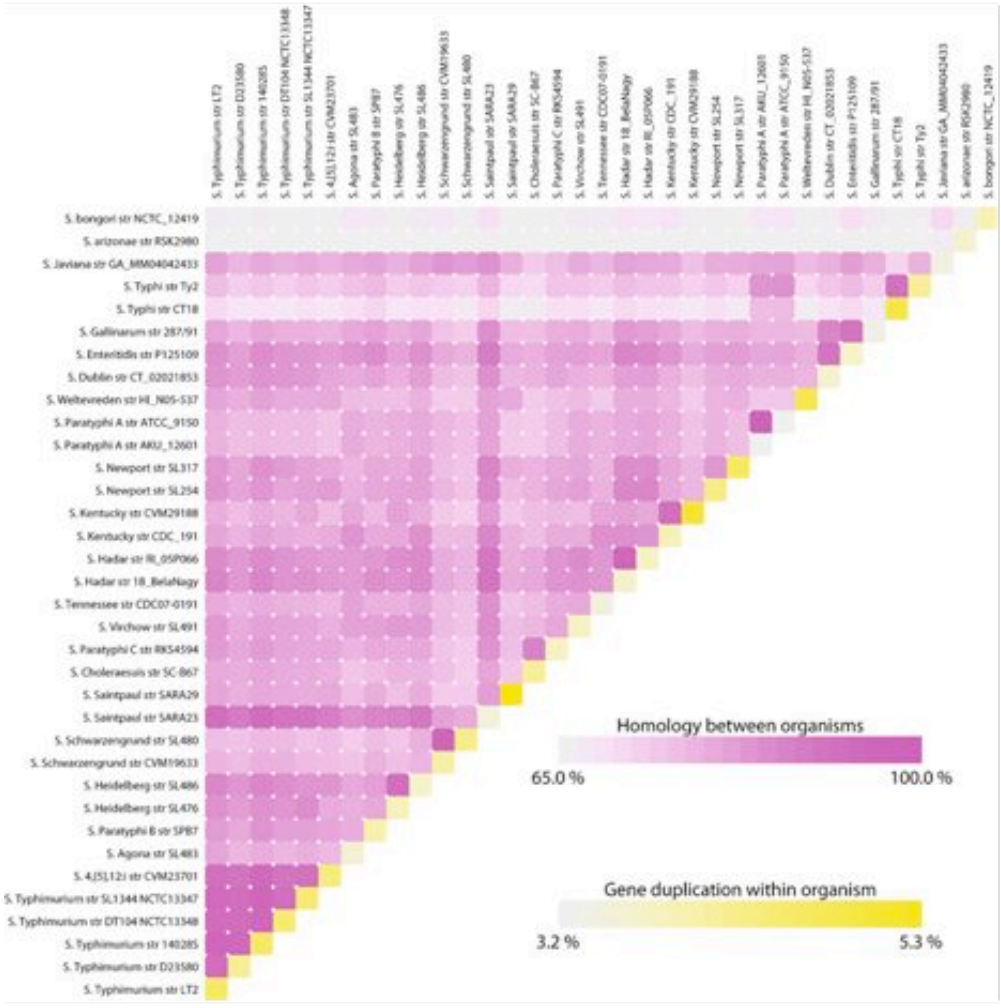


Figure 4 BLAST Matrix of 35 Salmonella genomes [68].

The visualization of gene conservation in a number of species against a reference genome was showed by BLAST atlas (Figure 5). Generally, the *Salmonella* strains were highly conserved

with most variation being concentrated in specific variable regions. The conservation of the *Salmonella* pathogenicity islands (SPIs) was typically high. In contrast to the host-specific serovar, *S*. Typhi str. Ty2 where SPI-7, the characteristic *S*. Typhi pathogenicity island [68]. SPI-7 is also found in both serovars *S*. Paratyphi C and *S*. Dublin [69,70]. SPI-1 and SPI-2 encode type III secretion systems (T3SS) [71]. T3SS of SPI-1 is responsible for the penetration of intestinal epithelium, whereas the T3SS of SPI-2 is important after access to macrophages [71]. SPI-1 is present in *S. bongori* and all subspecies and serotypes of *S. enterica* [72,73]. Previous study hypothesized that SPI-1 is a rather ancient acquisition gained at the separation of the genera *Escherichia* and *Salmonella* from a common ancestor [74]. The BLAST atlas (Figure 5) also revealed several genomic regions, which were absent from most or all *Salmonella* genomes. These regions are gene islands likely of viral origin for instance the region marked "I-a" is flanked by several genes encoding integrase/recombinase-like proteins and contains several phage-related proteins [68].

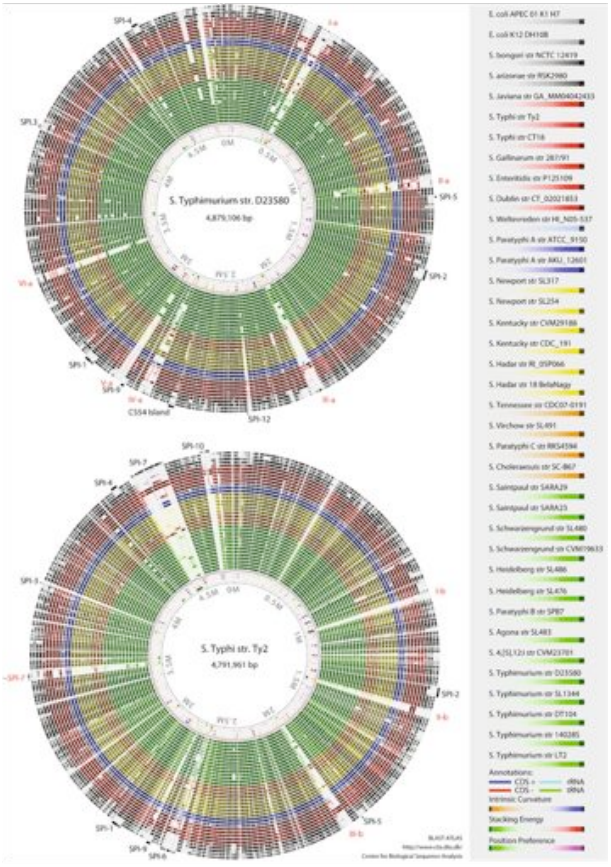


Figure 5 BLAST atlas of the 35 S. enterica and two E. coli. BLAST atlas with S. Typhimurium str. D23580 as reference (top). BLAST atlas with S. Typhi str. Ty2 as reference (bottom) [68].

Of all the 22 reported SPIs so far, only SPI-1, SPI-4 and SPI-9 are present in *S. bongori* [9]. SPI-3 exists as two independent insertions in *S. bongori*, SPI-3a and SPI-3b, that appear to have fused into a single element in *S. enterica* [75]. Another distinguishing feature of *S. bongori* is the lack of SPI-2 [72,73,76]. SPI-2 is composed of two distinct elements [77]. A portion of 25 kb is only present in *S. enterica*. This element is essential for systemic pathogenesis and encodes a second T3SS in *S. enterica* that is activated by intracellular bacteria. Another portion of SPI-2 is 15 kb element that was detected in *S. bongori* and *S. enterica*. This element is dispensable for

systemic virulence and encodes the tetrathionate reductase (Ttr) involved in anaerobic respiration [9,78].

Standardized procedures for identification of relevant genes and of variation are needed to enable comparison between studies and over time [**I**]. With the increasing number of available bacterial genome sequences (there were 73 publicly available *Salmonella* genomes in April 2011), when these genomes are compared, the genetic variation within bacterial species is greater than previously predicted [79,80]. In order to further investigate an outbreak caused by *Salmonella*, characterization of *Salmonella* isolates from genome data is a crucial step. *Salmonella* genomes are highly similar, particularly within subspecies *enterica*, where little variance exists in the genomes [68].

16S rRNA tree

The ribosomal genes are essential for the survival of all cells, and their structure cannot change much because of their involvement in protein synthesis [81]. Thus, 16S rRNA genes are highly conserved among isolates belonging to the same bacterial species [82]. Exceptions may be *N. meningitidis* [83] and *Mycoplasma* [84]. However, due to limited variation within a given species, the 16S sequencing is often not useful for epidemiological studies, where the classification of highly similar strains is needed. A phylogenetic tree based on 16S rRNA genes, extracted from 59 *Salmonella enterica* genomes [**I**], using RNAmmer [85] was shown in Figure 6A. As expected, there is not sufficient resolution to distinguish among the *Salmonella* subspecies *enterica* [**I**].

MLST tree

The limitations of using a single gene may be improved by the simultaneous analysis of multiple genes. Multi Locus Sequence Typing (MLST) has found wide applications, especially in phylogenetic studies. MLST tree is commonly based on seven housekeeping genes which each bacterial species have its own set. For *Salmonella*, these are: aroC, dnaN, hemD, hisD, purE, sucA and thrA (http://www.mlst.net). The MLST tree, based on an *in silico* analysis of the 73 available *Salmonella enterica* genomes, was shown in Figure 6B [**I**]. Strains of the same serovar

generally cluster into distinct groups, although exceptions exist; for example the *S.* Weltevreden str. HI_N05-537 was mixed with outbreak strains, *S.* Montivideo [**I**].

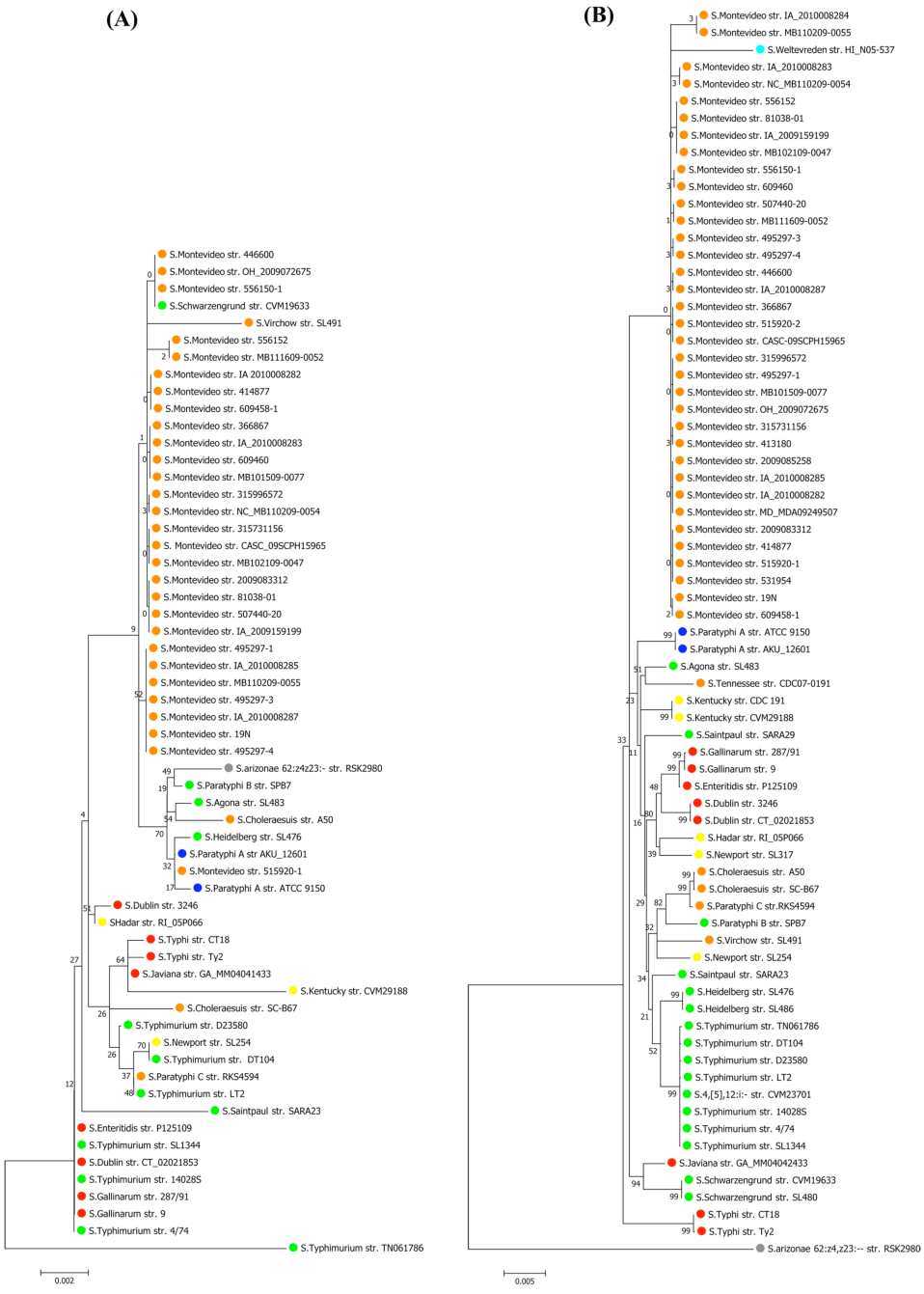


Figure 6 In silico 16S rRNA tree (A) and MLST tree (B). Colours represent a different serogroup (O antigen). The confidence value is the bootstrap value calculated by sampling with replacement from the multiple sequence alignment [**I**].

Furthermore, previous work on 61 sequenced *E.coli* genomes [82], found that the 16S rRNA tree cannot resolve well within the genus level and also that MLST cannot differentiate pathogenic strains from non- pathogenic strains. Still, MLST has proven useful for long-term analysis of population structures, but often fails to detect differences between closely related strains [45]. Indeed, improved MLST schemes that include more than 7 genes have been suggested [**I**].

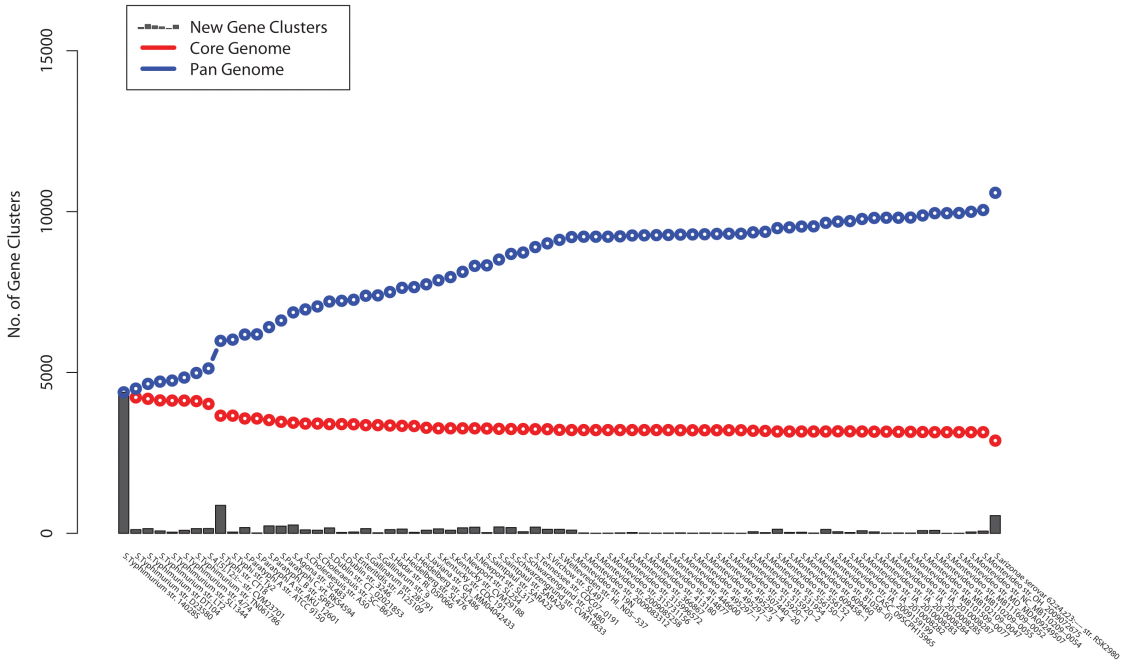
Salmonella enterica core genes

Sets of pan- and core-genomes were estimated, based on 73 *Salmonella* genomes [**I**] using a previously published method [86,87] which employs single-linkage clustering on top of BLASTp alignments. The progression of the pan- and core-genomes was shown in Figure 7A. The number of novel gene clusters in the pan-genome gradually increases when more genomes are considered, while the number of conserved gene clusters constituting the core genome decreases slightly. When all the *Salmonella* genomes had been considered, there were 10,581 pan gene clusters and 2,882 core gene clusters within species *enterica*. In the step going from *S.* Typhimurium to *S.* Typhi, the number of core genes dropped suddenly, most likely because the *S.* Typhi genome has undergone considerable pseudogene formation resulting in gene loss [88]. The number of core genes dropped again when adding a genome of the sub-species *arizonae* which is associated with cold-blooded animals. This technique has previously been applied successfully in finding core genomes for Proteobacteria genera *Burkholderia* [89], *Escherichia coli* [82], *Vibrionaceae* [90] and *Campylobacter jenuni* [87], as well as Bacteroides [91] and Lactic acid bacteria [92].

Genomic variation within the core genes

The core genes as calculated above were used for constructing a gene variation plot by performing all-against-all BLAST alignments between 2,882 core gene clusters and the 73 *Salmonella enterica* genomes. The resulting average identities within each core gene cluster was displayed in Figure 7B. From this figure, the average percent identity was very high (> 98%) in most of the core genes, but dropped sharply for around 5% of the core genes. The identified core genes can be divided into two categories: a small group of highly variable genes and the majority of genes, which showed little variation [**I**].

(A)



(B)

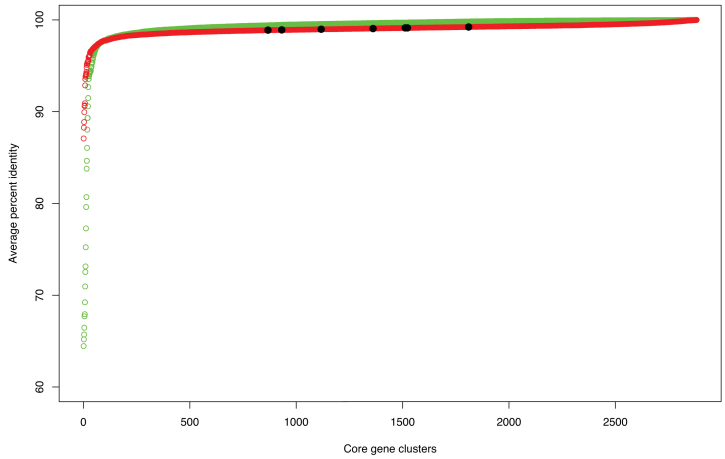


Figure 7 Pan- core-genome plot and variation plot. (A) Pan- and core-genome plot of 73 Salmonella enterica. The plot shows an increase of the pan-genome (blue line) and a decrease of the core-genome (red line) as more genomes are added. The last points show the total number of gene clusters in the pan-genome and the core-genome. (B) Variation plot. This plot shows the variation within core gene clusters in amino acid sequences (green dots) and nucleotide sequences (red dots). Black dots represent the distribution of housekeeping genes in the core genes. The Y- and X-axes represent average percent identity and numerical core gene cluster name respectively [**I**].

For the highly variable core genes, the variation in amino acid sequences (Figure 7B, green dots) was higher than for the nucleotide sequences (Figure 7B, red dots), whereas the opposite was the case for the more conserved core genes. This indicates that for core genes with low variation, there is a selection against mutations leading to amino acid changes, whereas for the highly variable genes, positive selection for amino acid changes seems to be the case. Therefore, the amino acid changes in highly variable core genes might be due to an increase in positive selection at some sites. Nonetheless, the importance of this needs to be confirmed by additional analysis [**I**].

The seven genes used for MLST were marked in the Figure 7B, and were scattered throughout the highly conserved part of the core genes (Figure 7B, black dots) and, as expected, little variation exists in these genes. Including core genes from both the highly conserved and variable regions might be beneficial in evolution studies. The more slowly evolving genes are useful in distinguishing between divergent and convergent evolution, while faster evolving genes can help in strain identification [**I**].

Consensus tree based on core genes

A total of 2,882 *Salmonella* core gene clusters were used for generating a consensus tree. Multiple alignments for each core gene cluster from all genomes were performed using MUSCLE [93]. A phylogenetic tree for each core gene was generated using PAUP [94]. The Phylip package was used to construct the consensus tree (Figure 8) from all the generated trees [95]

The tree generally divided the serotypes up well, but the bootstrap value in several branches was very low. This uncertainty could be due to the large number of core gene trees being analyzed individually; the low bootstrap values near the root reflect a lack of consensus at the higher levels. In contrast, the low bootstrap values found in *S.* Montevideo strains likely reflect uncertainty due to the high similarity of gene sequence of the clonal outbreak. All *S.* Montevideo strains sequenced were from a single outbreak [96] and as expected this analysis confirmed the almost complete identity of these isolates [**I**].

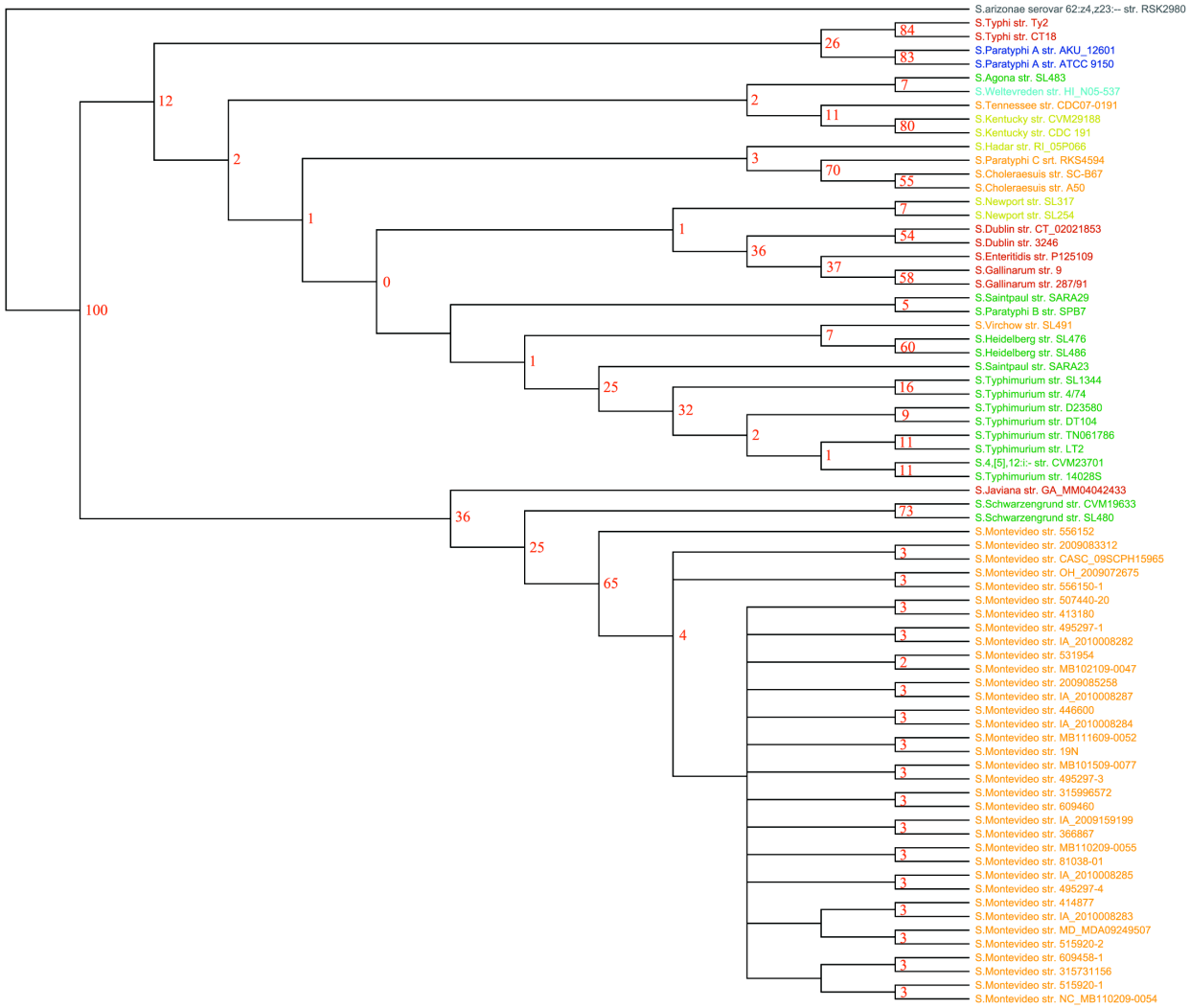


Figure 8 Consensus tree based on 2,882 core gene clusters. The percentage of branches present in all trees is shown. The colors represent different serogroups, as in Figure 6 [**I**].

Pan-genome tree

In principle, genome similarity is not only measurable by shared genes, but also by the absence of genes. Figure 9 is pan-genome tree, based on gene presence/absence across the analyzed *Salmonella* genomes [86]. The Pan-genome matrix consists of gene clusters (rows) and genomes (columns). The absence and presence of genes across genomes are represented by 0's and 1's respectively. The relative Manhattan distance between genomes was calculated and used for hierarchical clustering. The bootstrap values was calculated in order to represent the confidence of branches [86]. This tree bore a striking resemblance to the consensus tree based on core genes (Figure 8), although the bootstrap values were higher in many of the branches, especially near the

root. Of all methods investigated in this study, the pan-genome tree presented itself as the best solution for a tree that can resolve strain differences in a biologically meaningful way, even if it would be expected to correlate more with phenotype than phylogeny. It is, however, important to note that creating pan-genome trees requires higher quality sequencing data and assemblies than what are typically obtained using short reads from next generation sequencing methodologies [**I**].

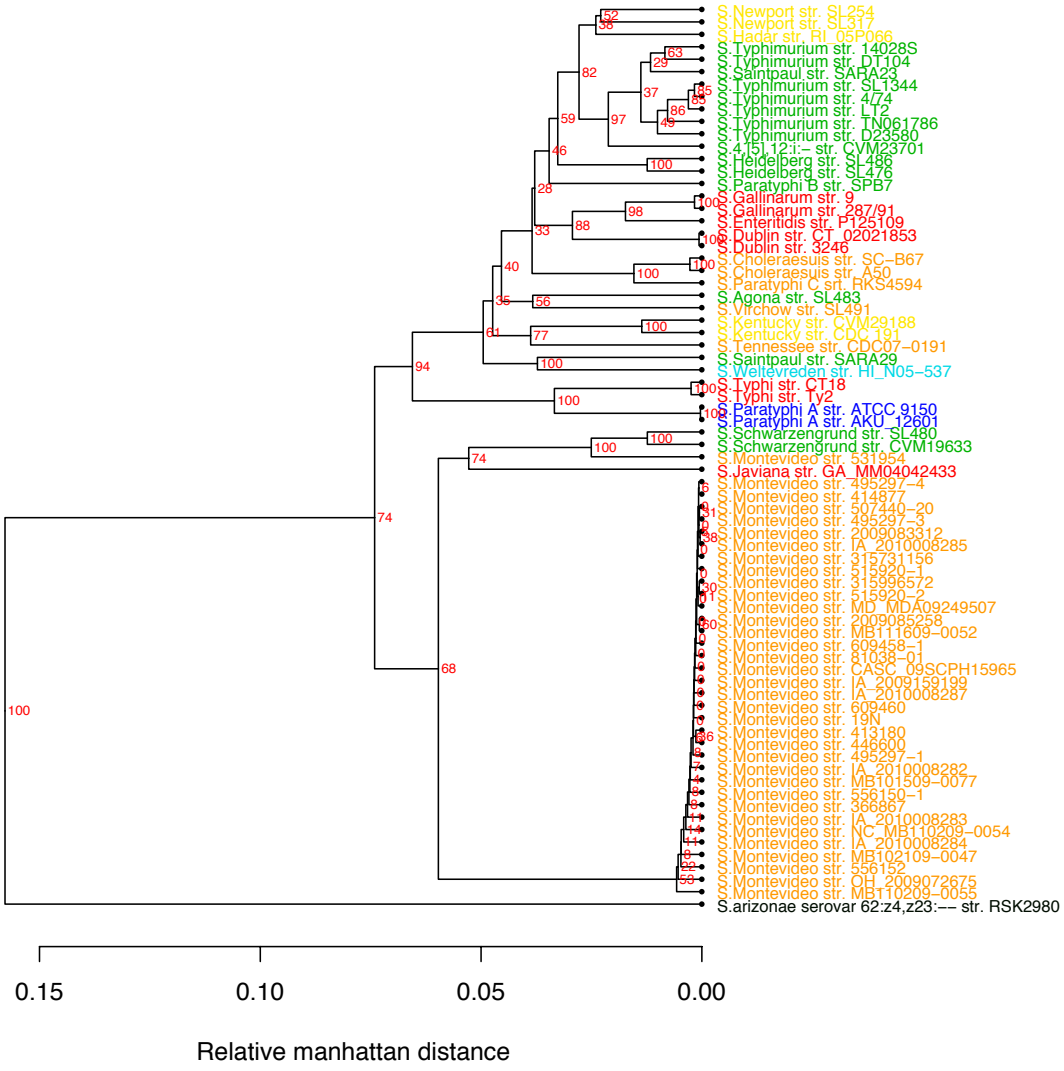


Figure 9 Pan-genome tree that is generated from the presence or absence of gene clusters across the Salmonella genomes. The bootstrap values are shown in red [**I**].

The power to discriminate between variants differs between the methods used. The phylogenetic analysis for the MLST tree is based on the identified informative sites among the seven

housekeeping genes, for the pan-genome tree on presence and absence of genes and for the consensus tree based on the informative sites of core gene clusters from alignments of all core genes. The number of informative sites for *in silico* MLST tree, pan-genome tree and consensus tree based on core gene clusters were 877 bp (10,008 total base-pairs in the seven genes), 7,699 genes (10,581 total genes) and 880,832 bp (2,868,821 bp in all core genes), respectively. The pan genome and core gene analysis were based on much more variation than the MLST analysis and have a much stronger power to discriminate closely related strains [**I**].

Bacterial typing should provide meaningful information for both epidemiological and evolutionary studies. For epidemiology, the ability to differentiate unrelated isolates (discriminatory power) and the ability to cluster related isolates are crucial. 16S rRNA and the MLST genes rarely provide separation between closely related strains. The performance of the pan-genome tree, however, is valid for epidemiological investigation in both discriminatory and clustering abilities. One caveat is that this method depends on good quality genomic data [**I**].

Comparative genomics can determine the conserved genes (core-genome) among bacterial genomes at either genus or species level. Genomic variation within the core-genome can then be used to reveal highly variable genes (fast evolving genes) and conserved genes (slow evolving genes). These core genes are useful for investigating molecular evolution and remain useful as candidate genes for bacterial genome typing–even if they cannot be expected to differentiate highly similar isolates from e.g. outbreak cases, such is not always desirable. Even in cases where a deeper distinction of isolates is of interest, e.g. in mapping outbreaks, core genes might still be useful as a reference fragment for SNPs calling instead of using whole genome analysis. However, in term of computational costs, the consensus tree based on core genes requires more computational time than the other methods [**I**].

WGS FOR OUTBREAK INVESTIGATION

An outbreak can be defined as a temporal increase in the incidence of infection (or colonisation) by a certain bacterial species, caused by enhanced transmission of a specific strain. It has to be noted that outbreaks can also be caused by multiple strains. The increased occurrence of a single strain, therefore, needs to be distinguished from the accumulation of sporadic cases [41]. Whole genome sequencing has been successfully used for elucidating the evolution and outbreak investigation of some *Salmonella* sub-types [96–100].

S. Montevideo outbreak

S. Montevideo is one of the top ten most common serovars associated with contaminated food in the US. Recently in the US, *S*. Montevideo was linked to the contamination in red and black pepper used in the production of Italian-style spiced meats in a New England processing facility causing a major salmonellosis outbreak that reportedly affected nearly 300 people in 44 states and the District of Columbia in 2009 and 2010 [96,98]. In a previous study [98], a total of 35 genomes of *S*. Montevideo collected from ingredient suppliers, patients and historically and geographically disparate food sources had been analyzed by PFGE and WGS. PFGE was unable to distinguish between outbreak and non-outbreak related strains whereas WGS based on SNP analysis was successful to resolve the outbreak [98]. WGS provided additional evidence that delimited the scope of the outbreak and suggested a domestic origin for *S*. Montevideo strain associated with this outbreak [96].

S. Enteritidis outbreak

According to the Centers for Disease Control and Prevention (CDC) in 2010, epidemiological investigations suggested that shell egg were the most likely source of a nationwide increase in *S*. Enteritidis isolates submitted to PulseNet (http://www.cdc.gov/salmonella/enteritidis/) [100] resulting that more than 500 million eggs involved during this nationwide were recalled [100].

In a previous study [100], a total of 106 *S*. Enteritidis isolates collected during the 2010 widespread shell egg contamination event within the Pulsed-Field Gel Electrophoresis (PFGE) pattern JEGX01.0004 and closet relatives were subjected to WGS.

SNP analysis revealed that the genetic diversity between different serovars included thousands of SNP differences whereas variability between the lineages of *S*. Enteritidis ranged only in the order of 100 to 600 SNP differences. The minimum number of SNP difference at 100 for an outbreak or clonal related strains is quite high. The isolates related to the 2010 egg shell outbreak clustered most closely together providing higher resolution for related strains within the contamination event and all the unrelated outbreak isolates retaining common PFGE patterns clustered outside the lineages of the 2010 egg outbreak. The result from WGS retrospectively supported the decision to recall a half a billion shell eggs by revealing SNP changes found in both eggs and hen houses and the changes were also shared with some food and clinical isolates [100].

These retrospective studies on outbreak investigation of *Salmonella* were conducted predominantly through SNP based phylogeny. However, prior to implementing WGS in routine surveillance and identification of outbreaks, reliable sub-typing criteria are essential. It is therefore essential to evaluate the WGS compared to traditional method and to determine which analytic approaches of WGS that might be most useful for a given bacterial species and sub-type.
A collection of 34 *S*. Typhimurium isolates was sequenced. This consisted of 18 isolates from 6 previously described outbreaks or clusters, primarily defined by MLVA [101,102] and 16 strains that were expected to be epidemiologically un-related to the outbreaks. The outbreaks were selected to cover outbreaks that were restricted in time and location [102] as well as some epidemiologically challenging outbreaks (outbreak 1–3) that lasted several months [101]. The isolates from each outbreak/cluster were selected to include some of the known diversity within these (e.g. based on phage type, MLVA, PFGE as well as the time span of the outbreak). The 16 background strains were selected, so at least two isolates belonged to the same phage type as that of each of the 6 outbreaks. In addition, 8 *S*. Enteritidis and 5 *S*. Derby were also sequenced and used for comparison [**II**].
A number of different bioinformatics approaches were applied on the data; including pan-genome tree, k-mer tree, nucleotide difference tree and SNP tree. The outcome was evaluated according to the pre-defined expected epidemiological data and also compared to results obtained using the conventional typing method, PFGE [**II**].

PFGE

Pulsed-field gel electrophoresis has been used as a standard procedure for epidemiological outbreak investigations of *Salmonella* [103]. Nonetheless, PFGE gave less discrimination power than WGS typing when applied to closely related strains, e.g. strains with the same phage type. Some strains from different outbreaks were grouped together and some outbreak strains were mixed with background isolates (Figure 10) [**II**]. PFGE is unable to separate very closely related strains because the low rate of genetic variation does not significantly impact the electrophoretic mobility of a restriction fragment [103,104].

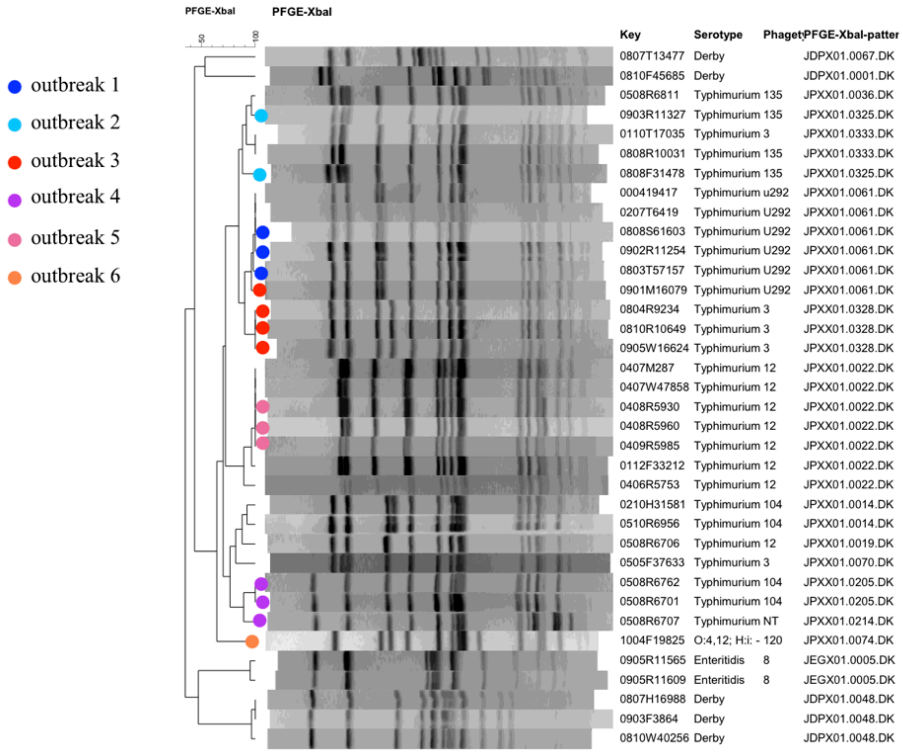


Figure 10 An UPGMA band based comparison of pulsed-field gel electrophoresis (PFGE) XbaI profiles [**II**].

Pan-genome tree

For the set of 34 *S.* Typhimurium genomes, pan-genome tree failed to cluster the outbreak strains into the corresponding groups of six different outbreak sources (Figure 11A). The tree only gave the reliable cluster for *S.* Derby outbreak strains (Figure 12A). Additionally, some different outbreak strains were mixed together. This method showed 65% and 64% concordance for the set of 34 and 47 genomes respectively. This is relatively low compared to the performance from other approaches (Table 1). However, the pan-genome tree revealed high performance for clustering strains according to their phage type [**II**].

Table 1 Evaluation results [**II**].

WGS typing methods	Percentage of concordance		Time (Minutes per genome)	Require reference	Type of input
	34 isolates	47 isolates			
Pan-genome tree	65	64	13	✖	Contigs
K-mer tree	88	89	5.2	✔	Contigs
Nucleotide difference tree	100	91	15	✔	Raw reads
SNP tree (raw reads)	100	91	20	✔	Raw reads
SNP tree (contigs)	100	89	5.5	✔	Contigs

K-mer tree

K-mer tree, alignment-free genome phylogeny, is constructed from the contiguous sequences of k bases called k-mers [105]. K can be any positive integer. In principle, sequences with high similarity likely share k-mers [106,107]. Based on this idea, the *de novo* assembled genomes were split into short sequences with the size of k (k-mers). K-mers were aligned against all the analyzed genomes. The number of hits or the frequency of k-mers across genomes was constructed as a matrix. The matrix consists of k-mers and genomes (rows and columns respectively) with the frequency of k-mers hits as a profile. The hierarchical clustering was performed in order to build the k-mer tree [**II**].

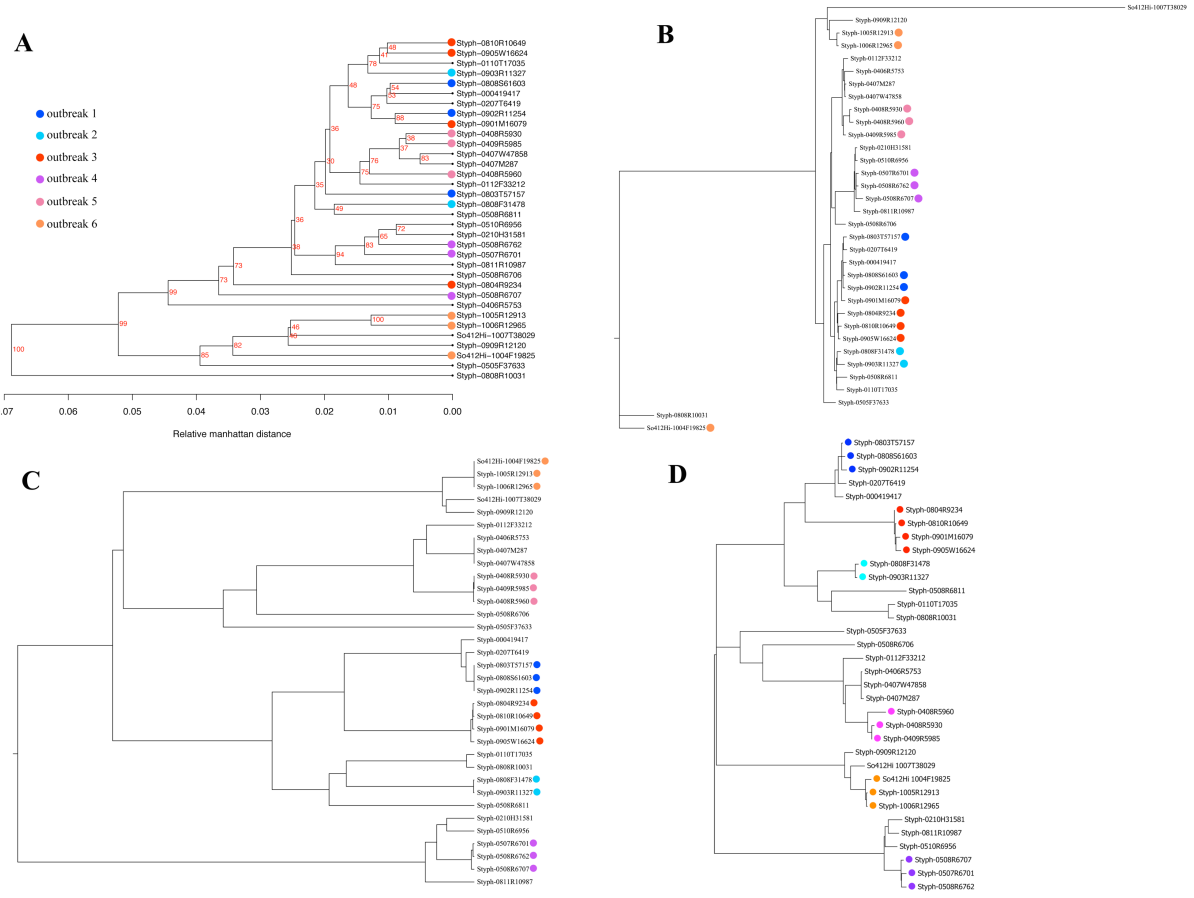


Figure 11 WGS typing results for the set of 34 genomes. (A) pan-genome tree, (B) k-mer tree, (C) nucleotide difference tree and (D) SNP tree. The test set consists of outbreak-related strains displayed with color label and non-related outbreak strains shown without coloring. The outbreak strains were labeled according to the six different outbreak sources [**II**].

K-mer tree gave higher resolution and more reliable tree than the pan-genome tree. However, some outbreak-related isolates were mixed up with the background strains (Figure 11B). Interestingly, the expanded tree in Figure 12B was capable to place the *S.* Enteritidis outbreak strains into two distinct clusters according to their outbreak groups. The tree also succeeded with clustering *S.* Derby outbreak strains suggested that the performance of k-mer tree remains unchanged when combining *Salmonella* strains from different serovars. This is most likely because the k-mer tree is independent from the reference genome. Nevertheless, the k-mer tree exhibited 88% and 89% concordance for the set of 34 and 47 isolates respectively (Table 1). An advantage of k-mer analysis is that the frequencies-based approach is much faster making the k-mer tree is the fastest method compared to the others [**II**].

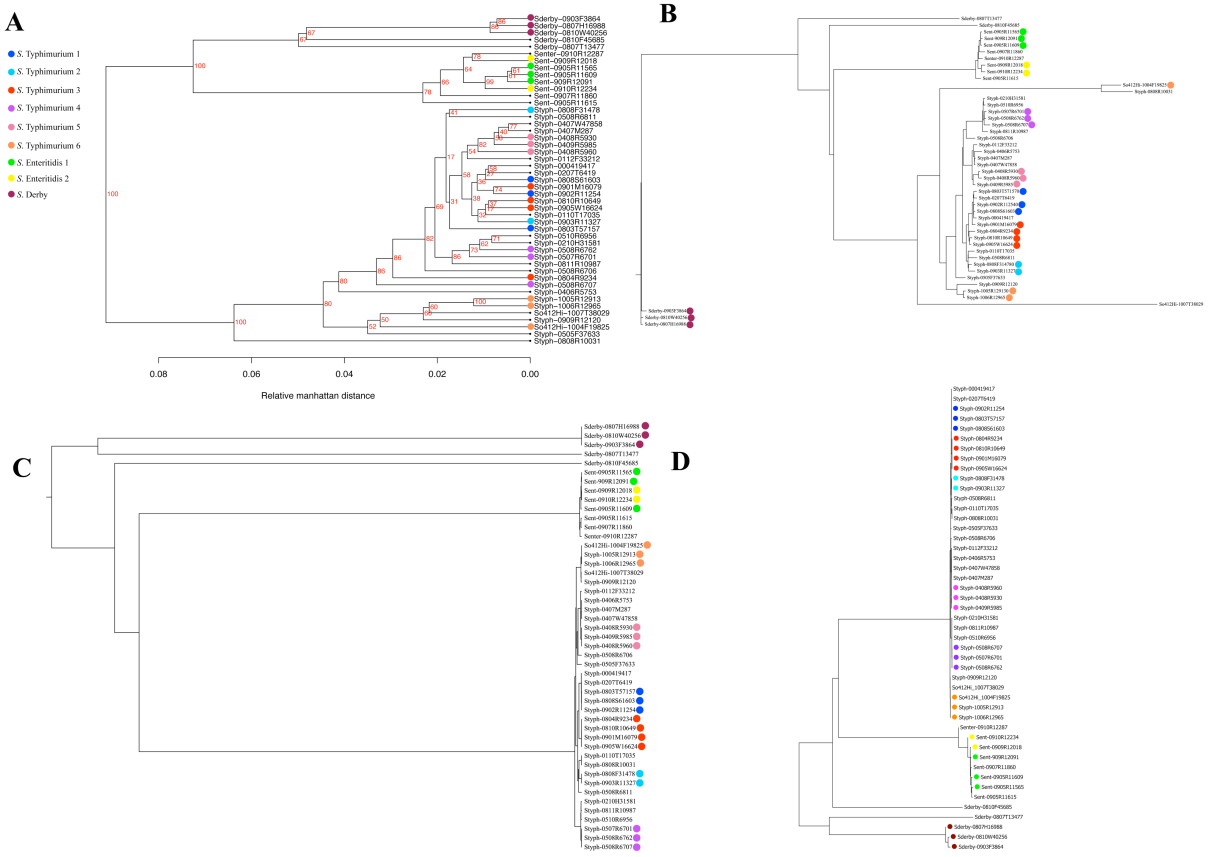


Figure 12 WGS typing results for the set of 47 genomes. (A) pan-genome tree, (B) k-mer tree, (C) nucleotide difference tree and (D) SNP tree. The labeled color is displayed the same as Figure 11 [**II**].

Nucleotide difference tree (ND tree)

The nucleotide difference tree (ND tree) is based on nucleotide difference between a pair of read mapped reference genomes. The well-studied *S*. Typhimurium str. LT2 was used as a reference genome (accession: AE006468, length of 4,857,432 bp). The reference genome was split into k-mers of length 17 and stored in a hash table. Each read with a length of at least 50 was split into 17-mers overlapping by 16. K-mers from the read and its reverse complement were mapped until an ungapped alignment with a score of at least 50 was found using a match score of 1 and a mismatch score of 23. When all reads had been mapped, each pair of sequences was compared and the number of nucleotide differences in positions called in all sequences was counted. A matrix with these numbers was given as input to a UPGMA algorithm implemented in the neighbor program (http://evolution.genetics.washington.edu/phylip.html) in order to construct the tree [**II**].

For the set of 34 *S*. Typhimurium, the ND tree classified outbreak-related strains into six obvious clusters (Figure 11C) with 100% concordance (Table 1). Thus, the typing ability of the ND tree was superior to the pan-genome tree and the k-mer tree. For the set of 47 genomes, the performance of the ND tree was slightly reduced (Figure 12C). The percentage of concordance decreased from 100 to 91% (Table 1) [**II**].

SNP tree

Single nucleotide polymorphisms (SNPs) were identified using a genobox pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [**III**]. The pipeline consists of various freely available programs. Basically, the paired-end reads from each isolates were aligned against the reference genome, *S*. Typhimurium str. LT2, using Burrows-Wheeler Aligner (BWA) [108]. SAMtools [109] 'mpileup' commands were used to determine and filter SNPs. The qualified SNPs were selected once they met the following criteria: (1) a minimum coverage (number of reads mapped to reference positions) of 20; (2) a minimum distance of 20 bps between each SNP; (3) a minimum quality score for each SNP at 30; and (4) all indels were excluded [**II**]. The qualified SNPs found within *Salmonella* core genes [**I**] were ultimately used to make SNP tree because SNPs within the noncore reflect the high proportion of mobile or extra-chromosomal elements, including prophage and genomic islands [110,111].

SNP tree was not only constructed from raw reads but also from contigs or assembled genomes. An application named Nucmer, which is a part of the software package called MUMmer version 3.23 [112] was introduced to align each of contigs to the reference genome. SNPs were determined from the resulting alignments using another MUMmer application called ‘‘show-snps’’ (with options ‘‘-CIlrT’’). The final set of SNPs was filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded [**II**].

For each genome, the final qualified SNPs were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script [**II**]. If a SNP is not found in the reference genome or the base coverage is less than a minimum setting (20 coverage), it is interpreted as not being a variation and the corresponding base in the reference is expected [104][**III**]. Subsequently, multiple alignments were employed by MUSCLE from MEGA5 [113]. SNP tree was constructed by MEGA5 using maximum parsimony method [113].

The SNP tree clustered *S*. Typhimurium outbreak-related strains into six clusters with 100% concordance (Table 1) and furthermore differentiated them accurately from the background isolates (Figure 11D). For the set of 47 genomes, SNP tree was able to categorized *S*. Derby isolates but unable to ultimately classify the *S*. Enteritidis strains (Figure 12D). The percentage of concordance was dropped from 100 to 91% (Table 1). This is due to the choice of reference genome, because this method depends heavily on the reference genome and this has to be closely related to the strains investigated for example the reference genome should be at least the same serovar as the strains under study. Using an inappropriate reference genome will cause exceed number of SNPs, which affects the final SNP tree [**II**].

In addition, SNP tree constructed from contigs exhibited slightly less concordance than the one from the raw reads (Table 1). In term of speed, the SNP tree from contigs can be achieved very fast (almost as fast as k-mer tree). It might be an alternative choice of using SNP tree for real-time typing. In addition, the identified SNPs were distributed thoroughly across core genes of the reference genome suggesting that the mutation occurred randomly through the core genes [**II**].

Figure 13 revealed that the minimum and maximum number of SNP difference within the outbreak strains were significantly less than those numbers between outbreak-related isolates and background isolates. The number of SNP difference between isolates within outbreaks ranged from 2 to 12 except the outbreak 5 (DT12) where the maximum number was relatively high (3–30 SNPs) suggesting that finding a general threshold to define an outbreak for *Salmonella* might

not be possible. Nonetheless, the SNP difference may be useful as an indicator of expected SNP distance in a particular serovar or a sub-outbreak cluster within serovar. Besides, the number of days within outbreak strains was unrelated to the number of SNP difference and this relation seems to be random [**II**].

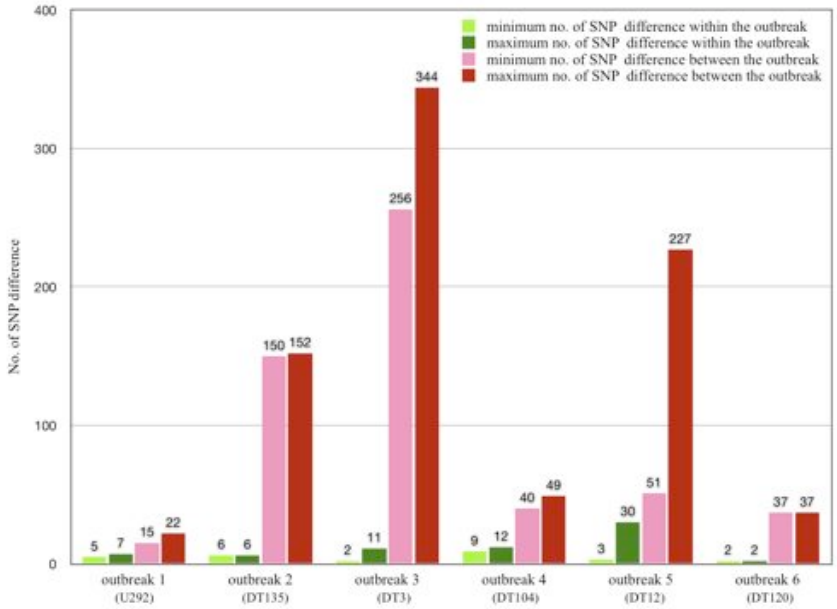


Figure 13 Minimum and maximum number of SNP difference. Green shaded bars show the minimum and maximum number of SNP difference between isolates within outbreaks and red shaded bars represent the number of SNP difference between outbreak-related isolates and background isolates [**II**].

Recent studies support SNP tree as an outbreak surveillance tool as mentioned [98,100,114,115]. Nonetheless, the SNP detection and validation need to be improved, and this method needs to be further evaluated in other bacterial pathogens to elucidate the usefulness of using SNP tree. Perhaps, for further pathogens, other approaches might be the more superior beside SNP analysis. In addition, it is especially a need to determine the importance of using different sequencing platforms, different analytic procedures and different reference strains for creating the SNP trees. Moreover, the robustness of this analytical approach for cluster detection in a routine setting has to be evaluated. The fact that the tree topology may give less resolution when new strains are added might cause some problems in the interpretation in a routine setting and over time [**II**].

This study suggests that WGS and analysis using SNP and/or nucleotide difference approaches are superior methodologies for epidemiological typing of *S.* Typhimurium isolates and might be very successfully applied for outbreak detection. For the very fast but rough result, k-mer tree might meet this requirement with constructing the tree in high speed and giving high accuracy in clade level [**II**].

It is also important to note that WGS is as all other typing tools to support for decision making and should always be used in combination with epidemiological and/or clinical information. For example, the different phylogenetic trees shown in this study were not meaningful without any support from epidemiological information. Thus, it is essential to combine epidemiological data and whole genome sequencing results [**II**].

snpTree SERVER

SNPs analysis has successfully been used in many recent studies on bacterial epidemiology and evolution [110,116,117]. Currently, There are a number of available non-commercial NGS genotype analysis software such as SOAP2 [118], GATK [119] and SAMtools [109]. Nonetheless, all of the software require bioinformatics skills, various settings and they do not have a user friendly web-interface.

The snpTree, a server for online-automatic SNP analysis and SNP tree construction from sequencing reads as well as from assembled genomes or contigs has been introduced. The server is a pipeline which integrates available SNP analysis software such as SAMtools [109] and MUMmer [112], with customized scripts. The performance of the server was evaluated using four published bacterial WGS data sets; *Vibrio cholerae* [120], *Staphylococcus aureus CC398* [117], *Salmonella Typhimurium* [121] and *Mycobacterium tuberculosis* [122]. The evaluation results for the first three cases were consistent and concordant for both raw reads and assembled genomes. In the latter case (*Mycobacterium tuberculosis*) the original publication involved extensive filtering of SNPs, which could not be repeated using snpTree [**III**].

The snpTree server might be not a perfect tool but it is an alternative choice for easy and rapid standardized and automatic SNP analysis tool in epidemiological studies. It is also useful for users with limited bioinformatics experience [**III**]. The web server is freely accessible at http://cge.cbs.dtu.dk/services/snpTree/.

Implementation of snpTree server

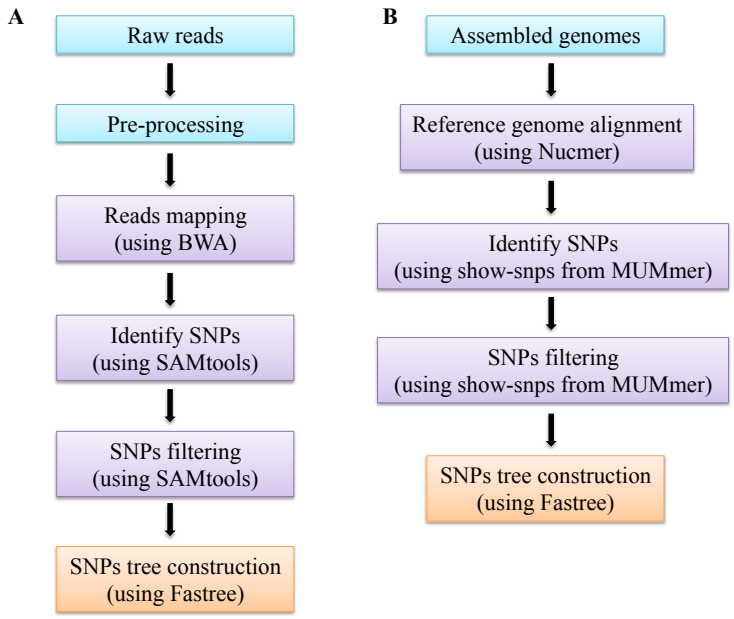


Figure 14 snpTree server implementation [**III**].

SNP tree construction from raw reads (Figure 14A), pre-processing (shown in blue) filters and trims raw data to remove low-quality bases. Trimmed raw reads are aligned against a reference genome by BWA [108] with mapping quality equal to 30 as a default. SNPs calling and filtering process (shown in purple) identifies and filters informative SNPs by SAMtools [109] with a couple of cut-offs, minimum coverage and minimum distance between each SNP (the default for both cut-offs is 10) and additionally all indels are filtered. SNPs tree construction step (shown in orange) transforms from multiple alignments of concatenated SNPs to a phylogenetic tree by using FasTree and a perl script. SNP tree construction from assembled genomes (Figure 14B), contigs or assembled genomes are aligned to a reference genome using Nucmer [112]. The SNPs calling and SNPs filtering steps are performed by a 'show-snps' application from MUMmer [112]. SNPs tree construction step is carried out as the same way as the raw reads [**III**].

snpTree server output

snpTree server provides an output to users with SNP tree figure in SVG format, number of SNPs and other relevant output files such as (i) SNP files, which contains identified SNPs including indels for each input genome in VCF format [123], (ii) concatenated SNPs in newick, phylip and fasta format, (iii) SNP annotation files giving an overview of nucleotide changes or amino acid changes from SNPs including SNPs containing input genomes as well as information about synonymous and non-synonymous SNPs [**III**]. An example of output is shown in Figure 15.

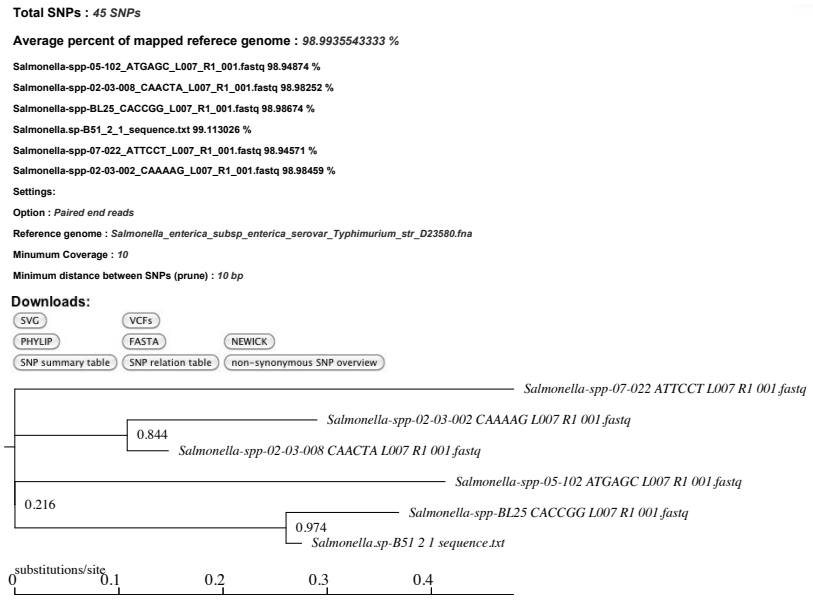


Figure 15 An example of the output from snpTree server using Illumina paired-end reads as input data [**III**].

WGS FOR GENOMIC EPIDEMIOLOGY

WGS cannot only be used for typing and outbreak investigation of *Salmonella*, but it also has been applied in epidemiology, population structure and evolutionary studies of some *Salmonella enterica* subtypes [97,124][**IV**].

Invasive S. Typhimurium in sub-Saharan Africa

However, severe infections with non-typhoidal *Salmonella* are relatively rare in Europe and North America, several studies have shown that invasive form of non-typhoidal *Salmonella* (iNTS) is endemic in many countries in sub-Saharan Africa [23,24,104]. The iNTS disease is common both in children with malnutrition, severe anemia, malaria or HIV and in infected adults [23]. The frequency of NTS- associated case fatalities can be extremely high in both adults and children (22–45%) [125]. A previous study of *S.* Typhimurium describing invasive diseases from 1997 to 2004, identified 31 isolates from Malawi and 13 out of 20 from Kenya to be of a novel multilocus sequence type (MLST) ST313 [24]. One *S.* Typhimurium ST313 isolate was sequenced and found to be phylogenetically distinct from other *S.* Typhimurium isolated in sub-Saharan Africa. It was suggested that *S.* Typhimurium ST313 is strongly associated with invasive

disease due to adaptation to human host as a result of genome degradation, similar to the evolutionary history of *S.* Typhi [24].

A retrospective study from 2012 [97] using WGS on a collection of 179 *Salmonella* Typhimurium isolates sampling between 1938 and 2010 from sub-Saharan Africa and different parts of the world showed that the lineage of sub-Saharan African isolates formed very tight clusters (with less SNP differences) and distinct from other *S.* Typhimurium found elsewhere in the world. A subset of 129 sub-Saharan invasive *S.* Typhimurium isolates from 7 sub-Saharan African countries sampling from 1988 to 2010 was further analyzed using BEAST (Bayesian Evolutionary Analysis Sampling Trees) [126,127] to reconstruct evolutionary history within the context of geographic distribution over time [97].

BEAST has been used widely in bacterial [128–131], viral [132,133] and eukaryotic [134] population studies. The mean evolutionary rates of the sub-Saharan African strains were estimated to be 1.9×10^{-7} to 3.9×10^{-7} substitutions per site per year. The rate is similar to the substitution rate in *Vibrio cholerae* (8×10^{-7} substitutions per site per year) [135] and resides between the estimated rates for *Yersinia pestis* (2×10^{-8}) [136] and *Staphylococcus aureus* (3×10^{-6}) [110]. Temporal phylogeny suggested that the most recent common ancestor of the *S.* Typhimurium was estimated to emerge ~52 years ago (95% highest posterior density (HPD) 1920.4–1979.5) and Malawi served as a potentially important earliest hub. In addition, the temporal emergence of the invasive *S.* Typhimurium also corresponds with the HIV pandemic in sub-Saharan Africa suggesting that the endemic of HIV might be one of many factors contributing the greater dissemination of invasive *S.* Typhimurium.

This study was one of the earliest studies of using WGS in spatial and temporal phylogeny for epidemiology and population structure of *Salmonella*. Besides, it provided the first whole genome based transmission study of the invasive *S.* Typhimurium from sub-Saharan Africa, and emphasized the power of WGS approaches to monitor the emergence and temporal spread of clonal bacterial populations associated with epidemics locally or globally [97].

Global genomic epidemiology of S. Typhimurium DT104

Globally, *Salmonella enterica* serovar Typhimurium is the most commonly isolated serovar [10]. During the last three decades, *S.* Typhimurium phage type DT104 emerged as the most important phage type and one of the best-studied because of its rapid global dissemination [10,137]. One of

the specialties of DT104 was its typically resistance to ampicillin, chloramphenicol, streptomycin, sulfonamide, and tetracycline (ACSSuT) [138] and its capacity to acquire extra resistance to other clinically important antimicrobial drugs [137]. Susceptible DT104 was first reported in 1960s, and subsequently as multidrug-resistant (MDR) DT104 in the early 1980s in the United Kingdom from humans and birds [139–141]. MDR DT104 rapidly emerged globally in 1990s and became the most prevalent reported phage type from humans and animals in many countries [137,139]. Previous epidemics with MDR phage types of *S*. Typhimurium, such as DTs 29, 204, 193 and 204c, were mostly restricted to cattle, whereas MDR DT104 spread among all domestic animals including cattle, poultry, pigs and sheep [139].

Despite several studies show that the origin and transmission routes of the phage type DT104 are still ambiguous. The transmission has been suggested to be through trade with live animals, but it has never been established whether the epidemiology in the different animal species are part of a common global spread or whether there are host specific variants [**IV**].

The recent study used WGS to study DT104 from mainly cattle and humans in Scotland sampled from 1990 to 2011 [124]. The study found relatively low animal-to-human or human-to-animal transitions and overall numbers of these transitions were similar suggesting that DT104 in Scotland circulated separately within each population with a low frequency of transferring in both directions and /or the animals and humans occurred in different and separate sources with also low level of transition. This study was severely hampered by the lack of inclusion of isolates from other animal species and by not including the fact that infections in humans are from food products of which most consumed in Scotland are imported from other countries [124]. Therefore, a carefully selected representative intercontinental DT104 collection from different sources in twenty-one countries covering the period from 1969 to 2012 were sequenced and subsequently analyzed based on temporally structured sequence analysis within a Bayesian framework aiming to exhibit population structure, phylogeny and evolution over time of DT104 as well as very recent disseminations events globally and locally between and within farms in Denmark [**IV**].

A total of 4,619 qualified SNPs were identified from all sampling 315 DT104 isolates. Phylogenomic dating was reconstructed using BEAST (Bayesian Evolutionary Analysis Sampling Trees) [126,127]. A combination of Bayesian Skyline model and relaxed uncorrelated lognormal clock were selected as population size change and molecular clock models.

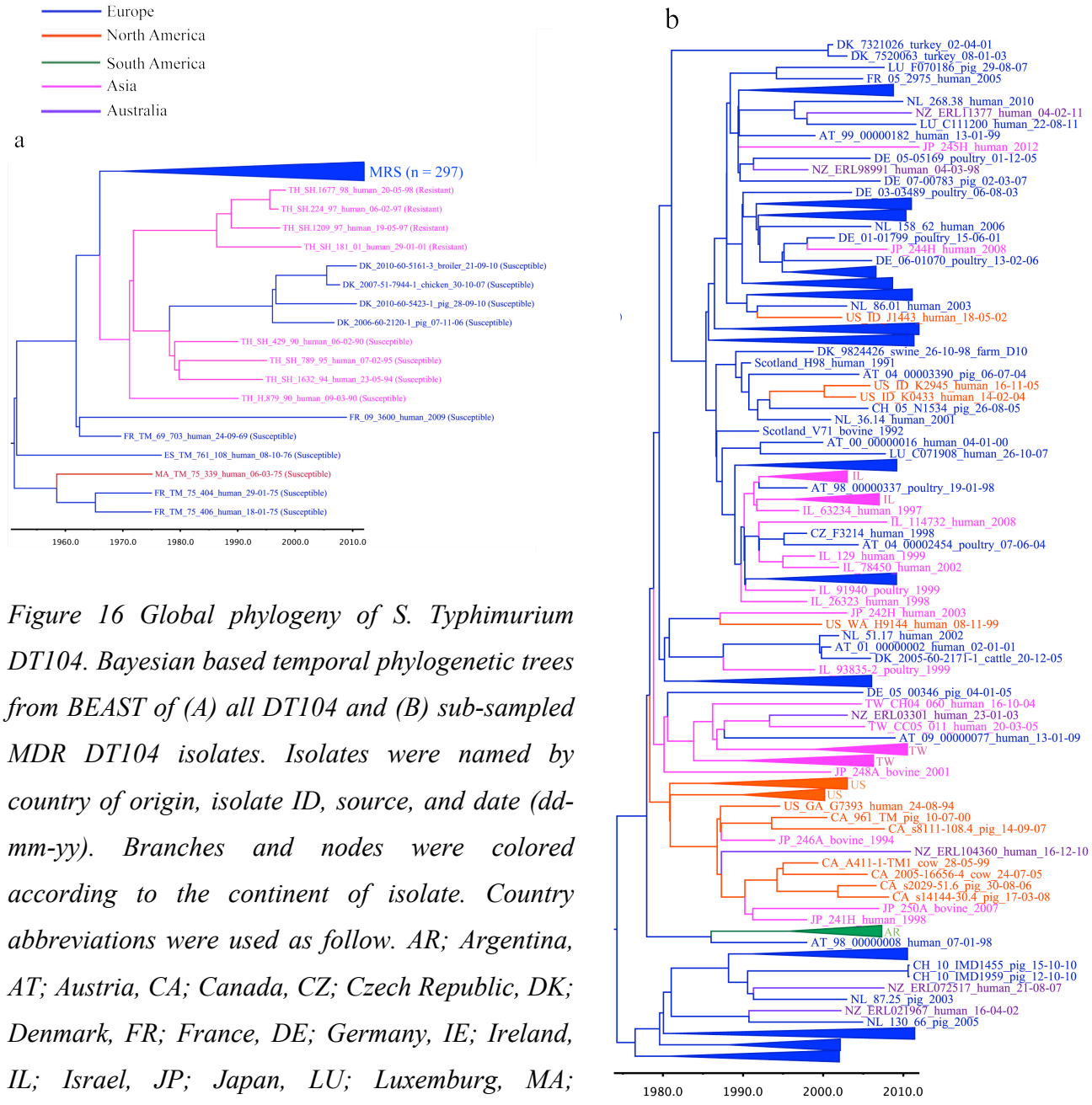


Figure 16 Global phylogeny of S. Typhimurium DT104. Bayesian based temporal phylogenetic trees from BEAST of (A) all DT104 and (B) sub-sampled MDR DT104 isolates. Isolates were named by country of origin, isolate ID, source, and date (dd-mm-yy). Branches and nodes were colored according to the continent of isolate. Country abbreviations were used as follow. AR; Argentina, AT; Austria, CA; Canada, CZ; Czech Republic, DK; Denmark, FR; France, DE; Germany, IE; Ireland, IL; Israel, JP; Japan, LU; Luxemburg, MA; Morocco, NL; The Netherlands, NZ; New Zealand, PL; Poland, ES; Spain, CH; Switzerland, TW; Taiwan; TH; Thailand, US; The United States [**IV**].

Bayesian based tree for all DT104 isolates was showed in Figure 16A. The mutation rate was estimated to be 2.97 x 10^{-7} SNP/site/year that was approximated to 1.47 SNP/year. The estimated rate of mutation corresponds to the mutation rates from previous studies of invasive *S.* Typhimurium in sub-Saharan Africa [97] and multidrug-resistant *S.* Typhimurium DT104 in

different hosts [124]. The most recent common ancestor was estimated to emerge in 1946 (95% highest posterior density, HPD, 1931 - 1959) as antimicrobial-susceptible DT104 in an unidentified reservoir. The earliest reports on susceptible DT104 strains isolated from human infections appeared in 1960s in the United Kingdom [139]. However, most if not all non-typhoidal *Salmonella* serovars have their natural reservoir in animals and only occasionally infect humans. Thus, susceptible DT104 may easily have spread for several years in an animal reservoir before the first infections occurred in humans.

The tree consisted of two individual clusters; a cluster of susceptible and resistant isolates and a complex cluster of multidrug-resistant strains with resistance to ampicillin, chloramphenicol, streptomycin, sulfonamide and tetracycline (ACSSuT resistance type). The susceptible and MDR clusters differed approximately by 109 SNPs. An average SNP difference among isolates in the susceptible cluster (n=18) was 103 SNPs, whereas that number among isolates in MDR cluster was only 60 SNPs (38 – 100 SNPs) despite a large number of isolates in the MDR cluster (n=297) suggesting that the MDR strains have higher degree of clonality [**IV**].

In contrast to the MDR strains, all of the isolates in the susceptible cluster contained small fragment or partial sequences of the 43-kb *Salmonella* genomic island 1 (SGI1, GenBank accession number AF261825) [142,143] and none of them harbored the 13-kb SGI1 multidrug resistance region [144]. The DT104 drug resistance genes can be transduced by P22-like phage ES18 and by phage PDT17, which are produced so far by all DT104 isolates [145]. The emergence of MDR strains would therefore cause by horizontal transfer of the DT104 antibiotic resistance gene cluster [146] into the SGI1-contained susceptible strains. The good evidence for horizontal transfer of the antibiotic resistance gene cluster is the presence of this cluster in another *S. enterica* serovar Agona [147]. This result challenges the hypothesis that the MDR DT104 emerged by acquiring an entire SGI1 with MDR region [146].

The 261 MDR isolates were analyzed separately yielding a total of 3,621 variable sites for Bayesian tree construction using BEAST (21B). The European isolates disseminated throughout the tree whereas the isolates from the other continents seem to be restricted to their continental origins except the human isolates from New Zealand that spread throughout the tree and clustered with isolates from different countries and continents (Figure 16B) suggesting that they might be travel-related cases. This result is concordant with the report that Australia and New Zealand have had few MDR DT104 human infections and most of human cases were from travellers

[137]. Another study found that 37% of Australian DT104 isolates were associated with travel aboard, especially to Southeast Asia [137].

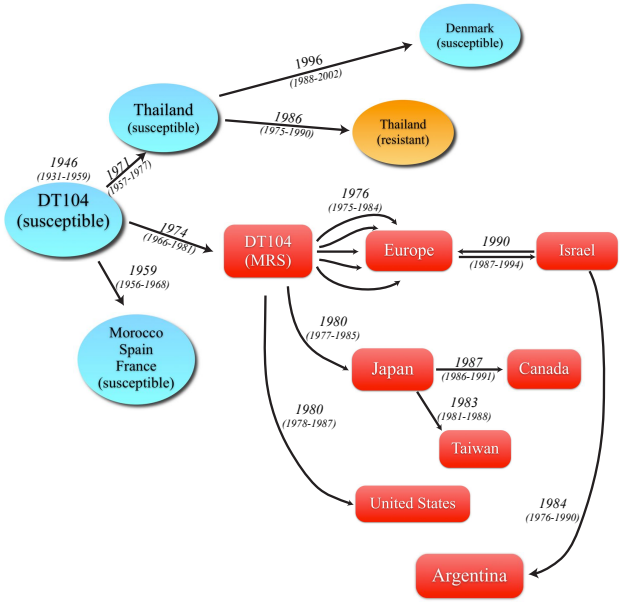


Figure 17 Diagram of the dissemination of S. Typhimurium DT104. Ages of nodes and divergence time of interested events from Figure 16A and 16B were summarized and illustrated in this diagram. Estimated time when transmission initially occurred (year) are represented as the median values, with 95% HPD in parenthesis [**IV**].

MDR DT104 was estimated to appear in ~1974 (95% HPD 1966 – 1981) (Figure 16B and Figure 17). From an unknown-source multiple introductions of MDR DT104 occurred to Europe from ~1976 (95% HPD 1975-1984). Subsequently another introduction to and from Israel occurred in ~1990 (95% HPD 1987-1994). Separated transmission routes occurred to Japan in ~1980 (95% HPD 1977-1985) and from Japan to Taiwan in ~1983 (95% HPD 1981-1988) and from Japan to Canada in ~1987 (95% HPD 1986-1991). In addition, the tree suggested that unknown-source MDR DT104 initially spread to the United States in ~1980 (95% HPD 1978-1987), consistent with the report of the emergence of MDR DT104 in the United States, particular in western states in early 1985 [148]. Furthermore, it spread from Israel to Argentina in ~1984 (95% HPD 1976-1990) with 81 average SNP difference (Figure 17).

Bayesian skyline plot for all DT104 isolates showed a demographic history of the DT104 from ~1960 (Figure 18). The effective population size of DT104 rose gradually until ~1980 after it became MDR DT104, and the population size increased sharply from 1980 to 1985 (Figure 18). This coincides with the estimated time of the occurrence of MDR DT104 in ~1974 (Figure 17) and the initial dissemination of MDR DT104 throughout Europe, Asia and America during 1980s (Figure 17).

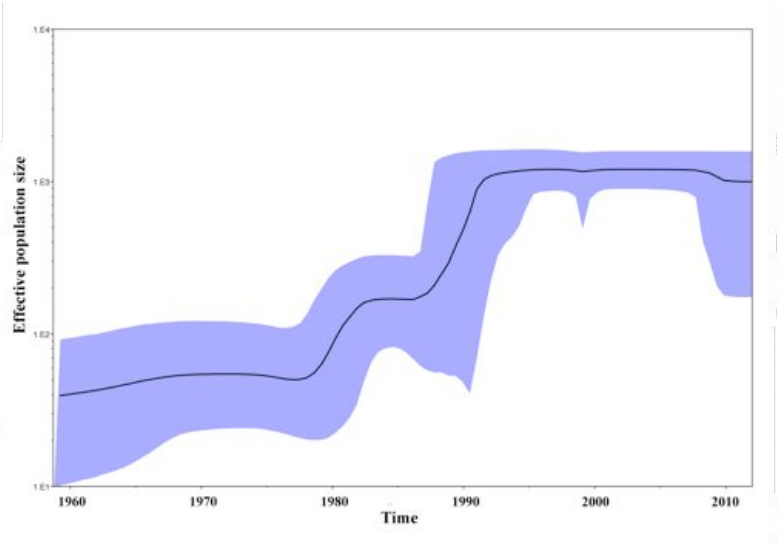


Figure 18 The changes in effective population size over time (year) of global DT104 [**IV**].

The second wave of DT104 started in ~1990, and the population size increased dramatically. This increasing may reflect the global dissemination of MDR DT104 because the timeline is agreeable with the occurrences of MDR DT104 in many countries. Germany had an increase in DT104 in the beginning of 1990s [149,150]. The number of DT104 human infections in UK rose from 259 in 1990 to 4006 in 1995 [151] as well as the number of DT104 in animals increased from 458 in 1993 to 1513 in 1996 [140]. Almost all 67% of *Salmonella* isolates from animals in Scotland during 1994-1995 were MDR DT104 [152], and a number of studies showed that throughout the 1990s, MDR DT104 spread to other parts of the world, including the United States, the United Kingdom, and France [35,148,153,154].The trend has leveled off since 1995 and gradually decreased from 2008 [**IV**].

Local genomic epidemiology of S. Typhimurium DT104 in Denmark

Seventy-five MDR *S.* Typhimurium DT104 isolates sampled from 1997 to 2011, originating from several farms in Denmark were sequenced. Sequence alignments of 755 SNPs were analyzed using BEAST. The Bayesian phylogenetic tree (Figure 19) established an estimated mutation rate at 2.15 x 10^{-7} SNP/site/year or 1.06 SNPs per year. The most recent common ancestor was predicted to emerge at the same period with the occurrence of the global MDR DT104 in ~1974 (95% HPD 1966 – 1981). The tree was divided into two major clusters and subsequently branched off to many lineages indicating multiple introductions of MDR DT104 to different farms in Denmark [**IV**].

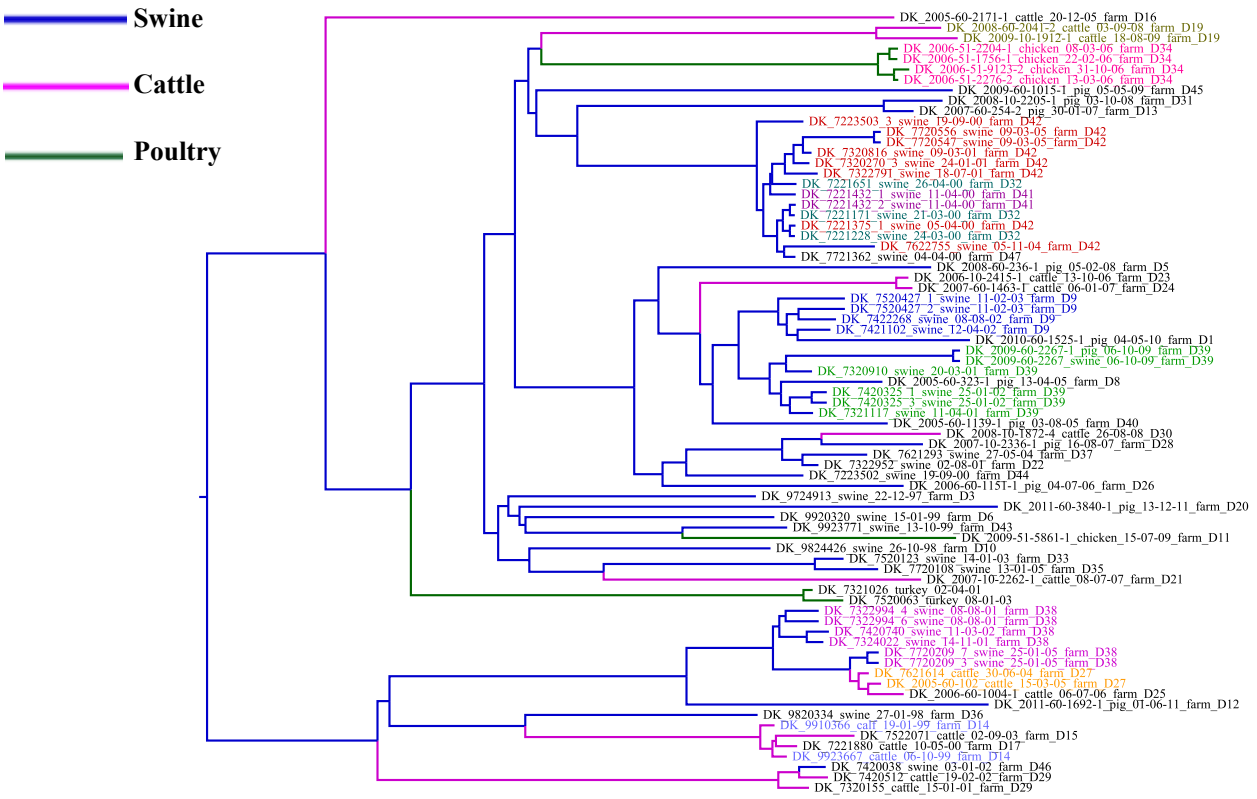


Figure 19 Local phylogeny of MDR S. Typhimurium DT104 isolates in Denmark. Farm numbers were noted at the end of node names. Nodes were colored according to farm of origin. A single isolate from a single farm was labeled in black. Colored branches showed animal sources [**IV**].

Several isolates were selected from the same farms. Most of those isolates were clustered phylogenetically according to their farms. Isolates from four different farms namely D32, D41, D42 and D47 were mixed into the same lineage. This is consistent with the information that there has been physical contact among those four farms, thus showing the ability of WGS to confirm very local epidemiology across animal herds. There were several branching links between isolates from swine and cattle (Figure 19), whereas isolates from poultry clustered separately. This indicates free transmission between cattle and swine, but a more closed spread in the poultry production. Concordantly, the analysis of proliferation of the infection in various species suggested that DT104 strains spread from cattle to pigs and humans [140][**IV**].

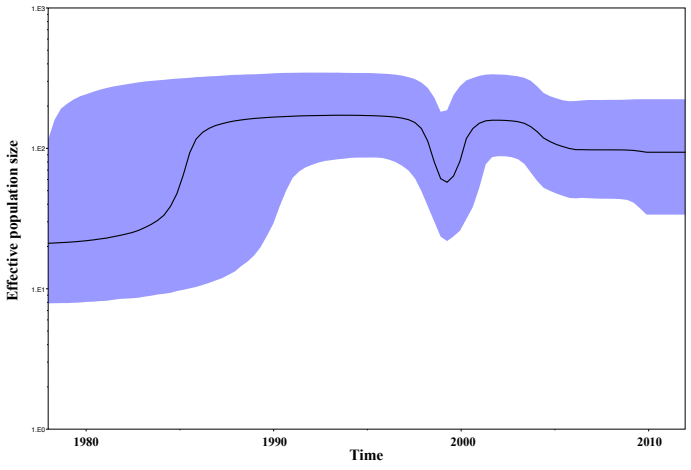


Figure 20 Bayesian skyline plot of changes in population size of Danish MDR DT104 over time [**IV**].

The relation between population structure and time (Figure 20) showed that the effective population size of MDR DT104 in Denmark rose slowly until ~1984 then it increased sharply from ~1984 to ~1987. Subsequently, the population was firmly established until ~1998 and it declined dramatically during ~1999 to ~2000, when an intensive eradication program was attempted in Denmark [155]. Following the abandon of the eradication program, the population size increased in ~2001 and decreased slightly from ~2004. Different Bayesian skyline plots based on sources were carried out. The pattern of sharp decline during 2000 has not been found among isolates from cattle, poultry and human except isolates from swine. In fact, 69% of Danish isolates were swine. Thus, the decline of the population size in 2000 was related to swine isolates. Therefore, the decreasing of swine MDR DT104 is an evidence of the accomplishment of the eradicating program in 1996 to 2000 established by the Federation of Danish Pig Producers and Slaughterhouse, in collaboration with the Danish Veterinary Service and the Danish Veterinary Laboratory. The program aimed to eradicate MDR DT104 from infected pig herds. The methods used included the depopulation of pig herds and the cleaning and disinfection of building before repopulation with pig free from DT104 [155].

Discrete phylogeographic analysis indicated several relationships among farms in Denmark. Average SNP distance between farms ranged from 3 to 100 SNPs. The confirmed contacts were concordant to the phylogeographic links showed in Figure 21. The contacts between farms D12-D38 and D41-D42 were direct relationships with 30 and 7 SNPs differences respectively, whereas the contacts from farms D32-D42 and D42-D47 were indirect contacts employed by 10 and 8 SNPs distances respectively [**IV**].

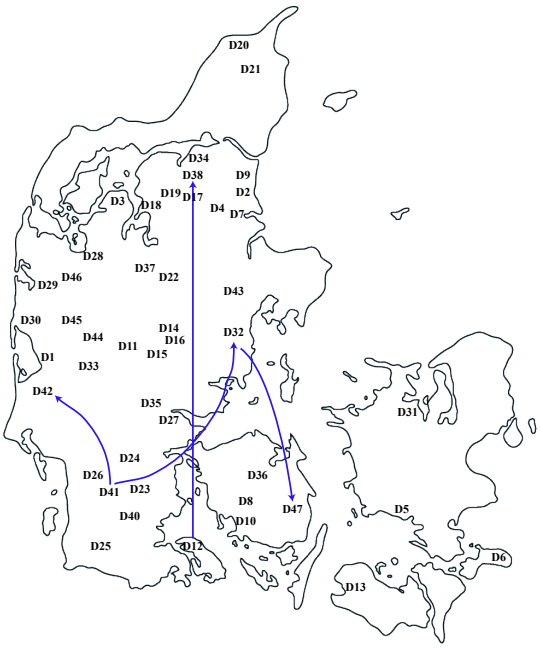


Figure 21 Confirmed geographic diffusion across different farms based on discrete phylogeographic analysis for the confirmed-farm contacts [**IV**].

This study shows the timeline of global and local disseminations of *S.* Typhimurium DT104 and the evolution of antimicrobial susceptible strains to MDR DT104 strains through horizontal transfer of 13-kb SGI MDR region. The results are consistent with many historical occurrences of MDR DT104 since it was observed in 1984. Moreover, the results carried out by WGS also confirm local epidemiology of DT104 and the efficiency of eradicating program in Denmark. The predicted transmission routes and demographic history would suggest any potential monitor and strategies for further prevention and control of similar successful clones [**IV**].

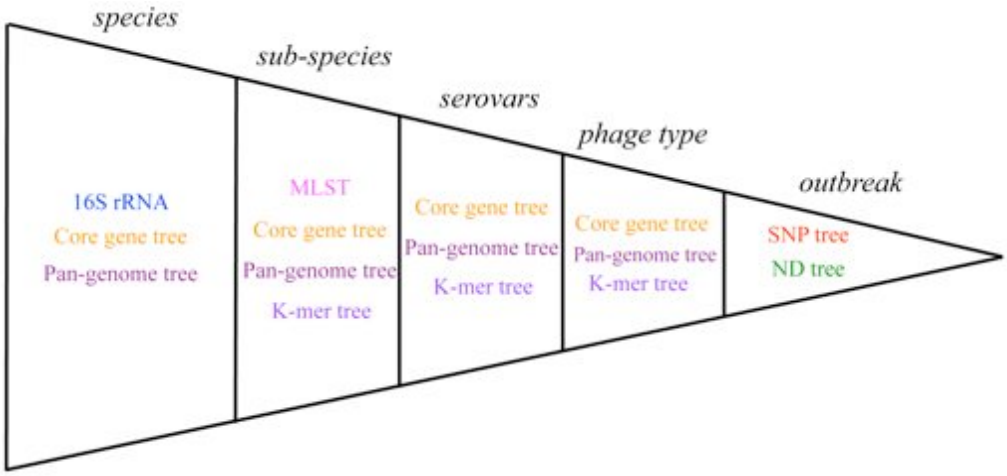


Figure 22 The performance of WGS approaches in each level of Salmonella typing proposed by the following studies in this thesis [**I, II, III, IV**].

In conclusion, the summarization of the application of WGS approaches, based on the studies included in this thesis [**I**, **II**, **III**, **IV**], for *Salmonella* typing in each typing level is illustrated in Figure 22.

FUTURE PREDICTIONS AND PERSPECTIVES

Global real-time surveillance and typing of *Salmonella* and other pathogens giving simultaneous information on bacterial typing and population structure, as well as outbreak detection are on the front line of incorporating whole genome sequencing for routine practice [156][**I,II**]. The advance of WGS and the use of epidemiological genomics emphasize the potential of practical application of WGS for clinical microbiology and underline the importance of developing reliable, fast and accurate genomics tools for clinical use [**III**].

Nonetheless, the current WGS approaches which is commonly relied on short reads, would be challenged by the novel upcoming single-molecule long-read sequencing technology such as Oxford Nanopore [60]. In addition, WGS techniques need to be improved to handle sequencing data across different sequencing platforms.

Sequencing without culturing or sequencing directly from the entire sample would be the next future of typing of pathogens [60]. There have been some studies using metagenomics approach to define the microbiomes of diverse samples and environments [157,158]. This is very useful particularly to overcome the low proportion of pathogen DNA in a clinical sample.

WGS is potentially useful for studying global and local epidemiology, population structure as well as short-term evolution of *S.* Typhimurium DT104 [**IV**]. Nevertheless, the longer term of evolution is also interesting for instance the evolutionary process from occupying one reservoir to another or from commensal to pathogen is still a puzzle. SNP tree approach may not be suitable for long-term evolution. Therefore, the genomic content as a target for long-term evolutionary studies need to be determined. In addition, virulence factor contributing to host specificity and host jump during long-term evolution is an interesting topic and might be revealed by phylogenetic modelling together with comparative genomics approaches.

CONCLUSIONS AND RECOMMENDATIONS

The studies included in this thesis have showed the advantages and the evaluation of using WGS for *Salmonella* typing. However, there is no a single methodology that universally used for all levels of typing. In epidemiology, the ability to differentiate unrelated isolates and the ability to cluster related isolates are crucial. 16S rRNA and the MLST genes rarely provide separation between closely related strains. Pan-genome and core genes are valid for sub-typing as such

serotype and phagetype. Meanwhile, SNP and ND approaches seem to be the most superior methodologies for epidemiological typing and outbreak investigation of *Salmonella*. Furthermore, SNP is also potentially useful as a tool for epidemiological study of global and local occurrence of *Salmonella*. However, it is important to note that WGS is as all other typing tools to support for decision-making and should always be used in combination with epidemiological and/or clinical information. Thus, it is essential to combine epidemiological data and whole genome sequencing results.

REFERENCES

1. Lan R, Reeves PR, Octavia S (2009) Population structure, origins and evolution of major Salmonella enterica clones. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 9: 996–1005.

2. Giannella R (1996) Salmonella: Epidemiology chapter 21. In: Baron S, editor. Baron's Medical Microbiology. University of Texas Medical Branch at Galveston.

3. Hohmann EL (2001) Nontyphoidal salmonellosis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 32: 263–269.

4. Grimont PAD, Grimont F, Bouvet P (2000) Taxonomy of the Genus Salmonella. In: Salmonella in domestic animals. Wray C, Wray A, editors Wallingford, United Kingdom: CABI Publishing.

5. Grimont PAD, Weill FX (2007) Antigenic formulae of the Salmonella serovars. 9th ed. Paris, France: WHO Collaborating Center for Reference and Research on Salmonella, Institut Pasteur.

6. Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B (2000) Salmonella nomenclature. J Clin Microbiol 38: 2465–2467.

7. Reeves MW, Evins GM, Heiba AA, Plikaytis BD, Farmer JJ (1989) Clonal nature of Salmonella typhi and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of Salmonella bongori comb. nov. Journal of clinical microbiology 27: 313–320.

8. Hohmann EL (2001) Nontyphoidal salmonellosis. Clin Infect Dis 32: 263–269.

9. Fookes M, Schroeder GN, Langridge GC, Blondel CJ, Mammina C, et al. (2011) Salmonella bongori provides insights into the evolution of the Salmonellae. PLoS pathogens 7: e1002191.

10. Lan R, Reeves PR, Octavia S (2009) Population structure, origins and evolution of major Salmonella enterica clones. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 9: 996–1005.

11. Wray C, Wray A (2000) Salmonella in Domestic Animals. Cabi Publishing, Wall- ingford.

12. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, et al. (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature 413: 852–856.

13. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, et al. (2008) Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome research 18: 1624–1637.

14. Uzzau S, Leori GS, Petruzzi V, Watson PR, Schianchi G, et al. (2001) Salmonella enterica serovar-host specificity does not correlate with the magnitude of intestinal invasion in sheep. Infection and immunity 69: 3092–3099.

15. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, et al. (2009) Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. BMC genomics 10: 36.

16. Jack EJ (1971) Salmonella abortion in sheep. Vet Annu 12: 57–63.

17. Taylor DN, Bied JM, Munro JS, Feldman RA (1982) Salmonella dublin infections in the United States, 1979-1980. The Journal of infectious diseases 146: 322–327.

18. Ikumapayi UN, Antonio M, Sonne-Hansen J, Biney E, Enwere G, et al. (2007) Molecular epidemiology of community-acquired invasive non-typhoidal Salmonella among children aged 2 29 months in rural Gambia and discovery of a new serovar, Salmonella enterica Dingiri. Journal of medical microbiology 56: 1479–1484.

19. Cohen JI, Bartlett JA, Corey GR (1987) Extra-intestinal manifestations of salmonella infections. Medicine 66: 349–388.

20. Sirichote P, Hasman H, Pulsrikarn C, Schønheyder HC, Samulioniené J, et al. (2010) Molecular characterization of extended-spectrum cephalosporinase-producing Salmonella enterica serovar Choleraesuis isolates from patients in Thailand and Denmark. Journal of clinical microbiology 48: 883–888.

21. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, et al. (2010) The global burden of nontyphoidal Salmonella gastroenteritis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 50: 882–889.

22. Brent AJ, Oundo JO, Mwangi I, Ochola L, Lowe B, et al. (2006) Salmonella bacteremia in Kenyan children. The Pediatric infectious disease journal 25: 230–236.

23. Graham SM (2010) Nontyphoidal salmonellosis in Africa. Current opinion in infectious diseases 23: 409–414.

24. Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, et al. (2009) Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. Genome research 19: 2279–2287.

25. Morpeth SC, Ramadhani HO, Crump JA (2009) Invasive non-Typhi Salmonella disease in Africa. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 49: 606–611.

26. Leekitcharoenphon P, Friis C, Zankari E, Svendsen CA, Price LB, et al. (2013) Genomics of an emerging clone of Salmonella serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo. Journal of infection in developing countries 7: 696–706.

27. Crump JA, Luby SP, Mintz ED (2004) The global burden of typhoid fever. Bulletin of the World Health Organization 82: 346–353.

28. Sirichote P, Hasman H, Pulsrikarn C, Schønheyder HC, Samulioniené J, et al. (2010) Molecular characterization of extended-spectrum cephalosporinase-producing Salmonella enterica serovar Choleraesuis isolates from patients in Thailand and Denmark. Journal of clinical microbiology 48: 883–888.

29. Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DMA, Jensen AB, et al. (2011) Global monitoring of Salmonella serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. Foodborne pathogens and disease 8: 887–900.

30. Olsen SJ, Bishop R, Brenner FW, Roels TH, Bean N, et al. (2001) The changing epidemiology of salmonella: trends in serotypes isolated from humans in the United States, 1987-1997. The Journal of infectious diseases 183: 753–761.

31. Xia S, Hendriksen RS, Xie Z, Huang L, Zhang J, et al. (2009) Molecular characterization and antimicrobial susceptibility of Salmonella isolates from infections in humans in Henan Province, China. Journal of clinical microbiology 47: 401–409.

32. Lauderdale T-L, Aarestrup FM, Chen P-C, Lai J-F, Wang H-Y, et al. (2006) Multidrug resistance among different serotypes of clinical Salmonella isolates in Taiwan. Diagnostic microbiology and infectious disease 55: 149–155.

33. Lee H-Y, Su L-H, Tsai M-H, Kim S-W, Chang H-H, et al. (2009) High rate of reduced susceptibility to ciprofloxacin and ceftriaxone among nontyphoid Salmonella clinical isolates in Asia. Antimicrobial agents and chemotherapy 53: 2696–2699.

34. Hendriksen RS, Bangtrakulnonth A, Pulsrikarn C, Pornruangwong S, Noppornphan G, et al. (2009) Risk factors and epidemiology of the ten most common Salmonella serovars from patients in Thailand: 2002-2007. Foodborne pathogens and disease 6: 1009–1019.

35. Rabsch W, Tschäpe H, Bäumler AJ (2001) Non-typhoidal salmonellosis: emerging problems. Microbes Infect 3: 237–247.

36. Hedberg C (1999) Food-related illness and death in the United States. Emerging infectious diseases 5: 840–842.

37. Aarestrup FM, Hendriksen RS, Lockett J, Gay K, Teates K, et al. (2007) International spread of multidrug-resistant Salmonella Schwarzengrund in food products. Emerging infectious diseases 13: 726–731.

38. Archambault M, Petrov P, Hendriksen RS, Asseva G, Bangtrakulnonth A, et al. (2006) Molecular characterization and occurrence of extended-spectrum beta-lactamase resistance genes among Salmonella enterica serovar Corvallis from Thailand, Bulgaria, and Denmark. Microbial drug resistance 12: 192–198.

39. Petrov P, Parmakova K, Siitonen A, Asseva G, Kauko T, et al. (2009) Salmonellosis cases caused by a rare Salmonella Enteritidis PT6c associated with travel to Bulgaria, June-July 2008. Euro surveillance 14.

40. Hendriksen RS, Bangtrakulnonth A, Pulsrikarn C, Pornreongwong S, Hasman H, et al. (2008) Antimicrobial resistance and molecular epidemiology of Salmonella Rissen from animals, food products, and patients in Thailand and Denmark. Foodborne pathogens and disease 5: 605–619.

41. Van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, et al. (2007) Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 13 Suppl 3: 1–46.

42. Miller JM (1993) Molecular technology for hospital epidemiology. Diagnostic microbiology and infectious disease 16: 153–157.

43. McQuiston JR, Parrenas R, Ortiz-Rivera M, Gheesling L, Brenner F, et al. (2004) Sequencing and comparative analysis of flagellin genes fliC, fljB, and flpA from Salmonella. Journal of clinical microbiology 42: 1923–1932.

44. Rabsch W (2007) Typhimurium Phage Typing for Pathogens. Methods Mol Biol 394: 177–211.

45. Li W, Raoult D, Fournier P-E (2009) Bacterial strain typing in the genomic era. FEMS microbiology reviews 33: 892–916.

46. Garaizar J, López-Molina N, Laconcha I, Lau Baggesen D, Rementeria A, et al. (2000) Suitability of PCR fingerprinting, infrequent-restriction-site PCR, and pulsed-field gel electrophoresis, combined with computerized gel analysis, in library typing of Salmonella enterica serovar enteritidis. Applied and environmental microbiology 66: 5273–5281.

47. Liebana E, Guns D, Garcia-Migura L, Woodward MJ, Clifton-Hadley FA, et al. (2001) Molecular typing of Salmonella serotypes prevalent in animals in England: assessment of methodology. Journal of clinical microbiology 39: 3609–3616.

48. Olsen JE, Skov MN, Threlfall EJ, Brown DJ (1994) Clonal lines of Salmonella enterica serotype Enteritidis documented by IS200-, ribo-, pulsed-field gel electrophoresis and RFLP typing. Journal of medical microbiology 40: 15–22.

49. Davis MA, Hancock DD, Besser TE, Call DR (2003) Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of Escherichia coli O157:H7. Journal of clinical microbiology 41: 1843–1849.

50. Lupski JR, Weinstock GM (1992) Short, interspersed repetitive DNA sequences in prokaryotic genomes. Journal of bacteriology 174: 4525–4529.

51. Van Belkum A, Scherer S, Van Alphen L, Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. Microbiology and molecular biology reviews : MMBR 62: 275–293.

52. Vergnaud G, Denoeud F (2000) Minisatellites: mutability and genome architecture. Genome research 10: 899–907.

53. Van Belkum A, Struelens M, De Visser A, Verbrugh H, Tibayrenc M (2001) Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clinical microbiology reviews 14: 547–560.

54. Ogata H, Audic S, Barbe V, Artiguenave F, Fournier PE, et al. (2000) Selfish DNA in protein-coding genes of Rickettsia. Science (New York, NY) 290: 347–350.

55. Fournier P-E, Zhu Y, Ogata H, Raoult D (2004) Use of highly variable intergenic spacer sequences for multispacer typing of Rickettsia conorii strains. Journal of clinical microbiology 42: 5757–5766.

56. Lindstedt B-A (2005) Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis 26: 2567–2582.

57. Van den Berg RJ, Schaap I, Templeton KE, Klaassen CHW, Kuijper EJ (2007) Typing and subtyping of Clostridium difficile isolates by using multiple-locus variable-number tandem-repeat analysis. Journal of clinical microbiology 45: 1024–1028.

58. Torpdahl M, Sørensen G, Lindstedt B-A, Nielsen EM (2007) Tandem repeat analysis for surveillance of human Salmonella Typhimurium infections. Emerging infectious diseases 13: 388–395.

59. Petersen RF, Litrup E, Larsson JT, Torpdahl M, Sørensen G, et al. (2011) Molecular Characterization of Salmonella Typhimurium Highly Successful Outbreak Strains. Foodborne Pathog Dis 8: 655–661.

60. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nature reviews Genetics 13: 601–612.

61. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology 30: 434–439.

62. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.

63. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475: 348–352.

64. Check Hayden E (2012) Nanopore genome sequencer makes its debut. Nature. doi:10.1038/ nature.2012.10051.

65. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, et al. (2014) Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. Journal of clinical microbiology 52: 139–146.

66. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, et al. (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS pathogens 8: e1002824.

67. Graham RMA, Doyle CJ, Jennison A V (2014) Real-time investigation of a Legionella pneumophila outbreak using whole genome sequencing. Epidemiology and infection: 1–5.

68. Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C (2011) The Salmonella enterica pan-genome. Microbial ecology 62: 487–504.

69. Liu W-Q, Liu G-R, Li J-Q, Xu G-M, Qi D, et al. (2007) Diverse genome structures of Salmonella paratyphi C. BMC genomics 8: 290.

70. Pickard D, Wain J, Baker S, Line A, Chohan S, et al. (2003) Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding Salmonella enterica pathogenicity island SPI-7. Journal of bacteriology 185: 5055–5065.

71. Hensel M (2004) Evolution of pathogenicity islands of Salmonella enterica. IJMM 294: 95–102.

72. Ochman H, Groisman EA (1996) Distribution of pathogenicity islands in Salmonella spp. Infection and immunity 64: 5410–5412.

73. Hensel M, Shea JE, Bäumler AJ, Gleeson C, Blattner F, et al. (1997) Analysis of the boundaries of Salmonella pathogenicity island 2 and the corresponding chromosomal region of Escherichia coli K-12. Journal of bacteriology 179: 1105–1111.

74. Bäumler AJ (1997) The record of horizontal gene transfer in Salmonella. Trends in microbiology 5: 318–322.

75. Blanc-Potard AB, Solomon F, Kayser J, Groisman EA (1999) The SPI-3 pathogenicity island of Salmonella enterica. Journal of bacteriology 181: 998–1004.

76. Knodler LA, Celli J, Hardt W-D, Vallance BA, Yip C, et al. (2002) Salmonella effectors within a single pathogenicity island are differentially expressed and translocated by separate type III secretion systems. Molecular microbiology 43: 1089–1103.

77. Hensel M, Nikolaus T, Egelseer C (1999) Molecular and functional analysis indicates a mosaic structure of Salmonella pathogenicity island 2. Molecular microbiology 31: 489–498.

78. Hensel M, Hinsley AP, Nikolaus T, Sawers G, Berks BC (1999) The genetic basis of tetrathionate respiration in Salmonella typhimurium. Molecular microbiology 32: 275–287.

79. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America 102: 13950–13955.

80. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, et al. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Functional & integrative genomics 6: 165–185.

81. Woese CR (1987) Bacterial evolution. Microbiological reviews 51: 221–271.

82. Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced Escherichia coli genomes. Microbial ecology 60: 708–720.

83. Sacchi CT, Whitney AM, Reeves MW, Mayer LW, Popovic T (2002) Sequence diversity of Neisseria meningitidis 16S rRNA genes and use of 16S rRNA gene sequencing as a molecular subtyping tool. Journal of clinical microbiology 40: 4520–4527.

84. Königsson MH, Bölske G, Johansson K-E (2002) Intraspecific variation in the 16S rRNA gene sequences of Mycoplasma agalactiae and Mycoplasma bovis strains. Veterinary microbiology 85: 209–220.

85. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research 35: 3100–3108.

86. Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. Standards in genomic sciences 2: 135–141.

87. Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, et al. (2010) Genomic characterization of Campylobacter jejuni strain M1. PloS one 5: e12253.

88. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, et al. (2009) Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. BMC genomics 10: 36.

89. Ussery DW, Kiil K, Lagesen K, Sicheritz-Pontén T, Bohlin J, et al. (2009) The genus burkholderia: analysis of 56 genomic sequences. Genome dynamics 6: 140–157.

90. Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, et al. (2010) On the origins of a Vibrio species. Microbial ecology 59: 1–13.

91. Karlsson FH, Ussery DW, Nielsen J, Nookaew I (2011) A closer look at bacteroides: phylogenetic relationship and genomic implications of a life in the human gut. Microbial ecology 61: 473–485.

92. Lukjancenko O, Ussery DW, Wassenaar TM (2012) Comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera. Microbial ecology 63: 651–673.

93. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5: 113.

94. Swofford D (2004) PAUP*. Phylogenetic Analysis Using Parsimony. Version 4 . Sinauer Associates.

95. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences Seattle: University of Washington.

96. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, et al. (2011) Identification of a salmonellosis outbreak by means of molecular sequencing. The New England journal of medicine 364: 981–982.

97. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, et al. (2012) Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. Nature genetics 44: 1215–1221.

98. Allard MW, Luo Y, Strain E, Li C, Keys CE, et al. (2012) High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC genomics 13: 32.

99. Zhou Z, McCann A, Litrup E, Murphy R, Cormican M, et al. (2013) Neutral genomic microevolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS genetics 9: e1003471.

100. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, et al. (2013) On the evolutionary history, population genetics and diversity among isolates of Salmonella Enteritidis PFGE pattern JEGX01.0004. PloS one 8: e55254.

101. Petersen RF, Litrup E, Larsson JT, Torpdahl M, Sørensen G, et al. (2011) Molecular characterization of Salmonella Typhimurium highly successful outbreak strains. Foodborne pathogens and disease 8: 655–661.

102. Torpdahl M, Sørensen G, Lindstedt B-A, Nielsen EM (2007) Tandem repeat analysis for surveillance of human Salmonella Typhimurium infections. Emerging infectious diseases 13: 388–395.

103. Foley SL, Zhao S, Walker RD (2007) Comparison of molecular typing methods for the differentiation of Salmonella foodborne pathogens. Foodborne pathogens and disease 4: 253–276.

104. Leekitcharoenphon P, Friis C, Zankari E, Svendsen CA, Price LB, et al. (2013) Genomics of an emerging clone of Salmonella serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo. Journal of infection in developing countries 7: 696–706.

105. Cheng J, Cao F, Liu Z (2013) AGP: a multimethods web server for alignment-free genome phylogeny. Molecular biology and evolution 30: 1032–1037.

106. DeSantis TZ, Keller K, Karaoz U, Alekseyenko A V, Singh NNS, et al. (2011) Simrank: Rapid and sensitive general-purpose k-mer search tool. BMC ecology 11: 11.

107. Yu H-J (2013) Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. Gene 518: 419–424.

108. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25: 1754–1760.

109. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25: 2078–2079.

110. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science (New York, NY) 327: 469–474.

111. Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, et al. (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS pathogens 7: e1002129.

112. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research 30: 2478–2483.

113. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution 28: 2731–2739.

114. Gieraltowski L, Julian E, Pringle J, Macdonald K, Quilliam D, et al. (2013) Nationwide outbreak of Salmonella Montevideo infections associated with contaminated imported black and red pepper: warehouse membership cards provide critical clues to identify the source. Epidemiology and infection 141: 1244–1252.

115. Hoffmann M, Luo Y, Lafon PC, Timme R, Allard MW, et al. (2013) Genome Sequences of Salmonella enterica Serovar Heidelberg Isolates Isolated in the United States from a Multistate Outbreak of Human Salmonella Infections. Genome announcements 1.

116. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, et al. (2012) Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proceedings of the National Academy of Sciences of the United States of America 109: 3065–3070.

117. Price LB, Stegger M, Hasman H, Aziz M, Larsen J, et al. (2012) Staphylococcus aureus CC398: host adaptation and emergence of methicillin resistance in livestock. mBio 3.

118. Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome research 19: 1124–1132.

119. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20: 1297–1303.

120. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, et al. (2011) Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. mBio 2: e00157–11.

121. Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, et al. (2012) High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal Salmonella typhimurium disease. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 54: 955–963.

122. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, et al. (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. The New England journal of medicine 364: 730–739.

123. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics (Oxford, England) 27: 2156–2158.

124. Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, et al. (2013) Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. Science 341: 1514–1517.

125. Gordon MA, Banda HT, Gondwe M, Gordon SB, Boeree MJ, et al. (2002) Non-typhoidal salmonella bacteraemia among HIV-infected Malawian adults: high mortality and frequent recrudescence. AIDS (London, England) 16: 1633–1641.

126. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 7: 214.

127. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution 29: 1969–1973.

128. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, et al. (2010) Evolutionary dynamics of Clostridium difficile over short and long time scales. Proceedings of the National Academy of Sciences of the United States of America 107: 7527–7532.

129. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. Science 331: 430–434.

130. Holt KE, Dolecek C, Chau TT, Duy PT, La TTP, et al. (2011) Temporal fluctuation of multidrug resistant salmonella typhi haplotypes in the mekong river delta region of Vietnam. PLoS neglected tropical diseases 5: e929.

131. Den Bakker HC, Bundrant BN, Fortes ED, Orsi RH, Wiedmann M (2010) A population genetics-based and phylogenetic approach to understanding the evolution of virulence in the genus Listeria. Applied and environmental microbiology 76: 6085–6100.

132. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122–1125.

133. Smith GJD, Bahl J, Vijaykrishna D, Zhang J, Poon LLM, et al. (2009) Dating the emergence of pandemic influenza viruses. Proceedings of the National Academy of Sciences of the United States of America 106: 11709–11712.

134. Endicott P, Ho SYW, Stringer C (2010) Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. Journal of human evolution 59: 87–95.

135. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, et al. (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477: 462–465.

136. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, et al. (2010) Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nature genetics 42: 1140–1143.

137. Helms M, Ethelberg S, Mølbak K (2005) International Salmonella Typhimurium DT104 infections, 1992-2001. Emerging infectious diseases 11: 859–867.

138. Mulvey MR, Boyd DA, Olson AB, Doublet B, Cloeckaert A (2006) The genetics of Salmonella genomic island 1. Microbes and infection / Institut Pasteur 8: 1915–1922.

139. Threlfall EJ (2000) Epidemic salmonella typhimurium DT 104--a truly international multiresistant clone. The Journal of antimicrobial chemotherapy 46: 7–10.

140. Poppe C, Smart N, Khakhria R, Johnson W, Spika J, et al. (1998) Salmonella typhimurium DT104: a virulent and drug-resistant pathogen. Can Vet J 39: 559–565.

141. Hollinger K, Wray C, Evans S, Pascoe S, Chappell S, et al. (1998) Salmonella Typhimurium DT104 in cattle in Great Britain. J Am Vet Med Assoc 213: 1732–1733.

142. Boyd DA, Peters GA, Ng L, Mulvey MR (2000) Partial characterization of a genomic island associated with the multidrug resistance region of Salmonella enterica Typhymurium DT104. FEMS microbiology letters 189: 285–291.

143. Hall RM (2010) Salmonella genomic islands and antibiotic resistance in Salmonella enterica. Future microbiology 5: 1525–1538.

144. Targant H, Doublet B, Aarestrup FM, Cloeckaert A, Madec J-Y (2010) IS6100-mediated genetic rearrangement within the complex class 1 integron In104 of the Salmonella genomic island 1. The Journal of antimicrobial chemotherapy 65: 1543–1545.

145. Schmieger H, Schicklmaier P (1999) Transduction of multiple drug resistance of Salmonella enterica serovar typhimurium DT104. FEMS microbiology letters 170: 251–256.

146. Cloeckaert A, Schwarz S (2001) Molecular characterization, spread and evolution of multidrug resistance in Salmonella enterica typhimurium DT104. Veterinary research 32: 301–310.

147. Cloeckaert A, Sidi Boumedine K, Flaujac G, Imberechts H, D'Hooghe I, et al. (2000) Occurrence of a Salmonella enterica serovar typhimurium DT104-like antibiotic resistance gene cluster including the floR gene in S. enterica serovar agona. Antimicrobial agents and chemotherapy 44: 1359–1361.

148. Glynn MK, Bopp C, Dewitt W, Dabney P, Mokhtar M, et al. (1998) Emergence of multidrug-resistant Salmonella enterica serotype typhimurium DT104 infections in the United States. The New England journal of medicine 338: 1333–1338.

149. Threlfall EJ, Ward LR, Frost JA, Willshaw GA (2000) Spread of resistance from food animals to man--the UK experience. Acta veterinaria Scandinavica Supplementum 93: 63–8; discussion 68–74.

150. Prager R, Liesegang A, Rabsch W, Gericke B, Thiel W, et al. (1999) Clonal relationship of Salmonella enterica serovar typhimurium phage type DT104 in Germany and Austria. Zentralblatt für Bakteriologie : international journal of medical microbiology 289: 399–414.

151. Threlfall EJ, Ward LR, Rowe B (1997) Increasing incidence of resistance to trimethoprim and ciprofloxacin in epidemic Salmonella typhimurium DT104 in England and Wales. Euro surveillance 2: 81–84.

152. Low JC, Angus M, Hopkins G, Munro D, Rankin SC (1997) Antimicrobial resistance of Salmonella enterica typhimurium DT104 isolates and investigation of strains with transferable apramycin resistance. Epidemiology and infection 118: 97–103.

153. Ward LR, Threlfall EJ, Rowe B (1990) Multiple drug resistance in salmonellae in England and Wales: a comparison between 1981 and 1988. Journal of clinical pathology 43: 563–566.

154. Witte W (1998) Medical consequences of antibiotic use in agriculture. Science 279: 996–997.

155. Baggesen DL, Aarestrup FM (1998) Characterisation of recently emerged multiple antibiotic-resistant Salmonella enterica serovar typhimurium DT104 and other multiresistant phage types from Danish pig herds. The Veterinary record 143: 95–97.

156. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, et al. (2014) Evaluation of Real-Time WGS for Routine Typing, Surveillance and Outbreak Detection of Verotoxigenic Escherichia coli. Journal of clinical microbiology.

157. Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nature reviews Genetics 13: 260–270.

158. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, et al. (2012) Experimental and analytical tools for studying the human microbiome. Nature reviews Genetics 13: 47–58.

Article I

Genomic variation in *Salmonella enterica* core genes for epidemiological typing.

ARTICLE I

Pimlapas Leekitcharoenphon,[1] Oksana Lukjancenko,[2] Carsten Friis,[1] Frank M. Aarestrup,[1] David W Ussery,[2*]

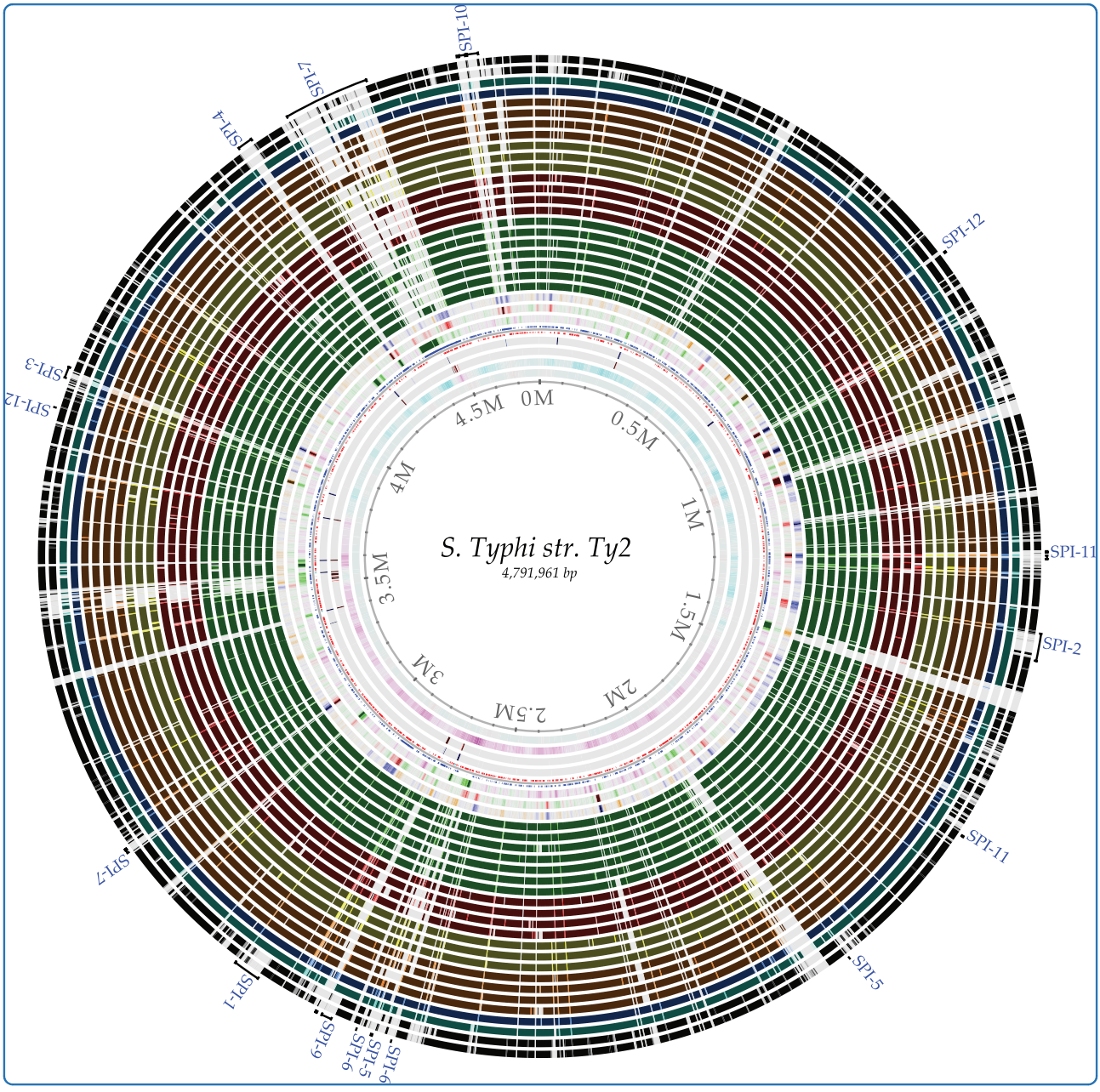
[1] National Food Institute, Building 204, Technical University of Denmark, 2800 Kgs Lyngby, Denmark

[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kgs Lyngby, Denmark

*Corresponding author: **David W Ussery,**
Center for Biological Sequence Analysis,
Department of Systems Biology,
Technical University of Denmark,
Building 208, DK-2800 Kgs Lyngby, Denmark
E-mail: dave@cbs.dtu.dk

BMC Genomics 2012 Mar 12;13:88.

BMC Genomics



Genomic variation in *Salmonella enterica* core genes for epidemiological typing

Leekitcharoenphon *et al.*

BioMed Central

Leekitcharoenphon *et al. BMC Genomics* 2012, **13**:88
http://www.biomedcentral.com/1471-2164/13/88 (12 March 2012)

BMC Genomics

RESEARCH ARTICLE Open Access

Genomic variation in *Salmonella enterica* core genes for epidemiological typing

Pimlapas Leekitcharoenphon[1,2], Oksana Lukjancenko[2], Carsten Friis[1], Frank M Aarestrup[1] and David W Ussery[2*]

Abstract

Background: Technological advances in high throughput genome sequencing are making whole genome sequencing (WGS) available as a routine tool for bacterial typing. Standardized procedures for identification of relevant genes and of variation are needed to enable comparison between studies and over time. The core genes–the genes that are conserved in all (or most) members of a genus or species–are potentially good candidates for investigating genomic variation in phylogeny and epidemiology.

Results: We identify a set of 2,882 core genes clusters based on 73 publicly available *Salmonella enterica* genomes and evaluate their value as typing targets, comparing whole genome typing and traditional methods such as 16S and MLST. A consensus tree based on variation of core genes gives much better resolution than 16S and MLST; the pan-genome family tree is similar to the consensus tree, but with higher confidence. The core genes can be divided into two categories: a few highly variable genes and a larger set of conserved core genes, with low variance. For the most variable core genes, the variance in amino acid sequences is higher than for the corresponding nucleotide sequences, suggesting that there is a positive selection towards mutations leading to amino acid changes.

Conclusions: Genomic variation within the core genome is useful for investigating molecular evolution and providing candidate genes for bacterial genome typing. Identification of genes with different degrees of variation is important especially in trend analysis.

Background

With the increasing number of available bacterial genome sequences, when these genomes are compared, the genetic variation within bacterial species is greater than previously predicted [1,2]. Rapid and reliable sub-typing of bacterial pathogens is important for identification of outbreaks and monitoring of trends in order to establish population structure and to study the evolution among bacterial genomes especially within and between the outbreak strains. Today, the most widely used typing methods for bacterial genomes include multilocus sequence typing (MLST), pulsed field gel electrophoresis (PFGE), sequencing of 16S rRNA genes, and multilocus variable-number of tandem-repeat analysis (MLVA).

PFGE and MLVA have major benefits, but are time consuming and the results are difficult to standardize [3]. Other typing methods which rely on one or a few ubiquitous genes, such as the 16S rRNA gene or a set of housekeeping genes in MLST, are capable of classification at the species level and sometimes also at the subspecies level, but the biological information in a narrow selection of genes will rarely be sufficient to clearly distinguish between closely related strains such as several isolates of the same serotype [4-6]. Thus, more of the genome content should be considered rather than just one or a few genes [4].

The price and time for whole genome sequencing will soon be in the same range as the traditional typing methods mentioned above. Genome sequencing can be a powerful method in epidemiological and evolutionary investigations [7-9]. Although, to date, this has only been used in more limited epidemiological investigations where isolates suspected to be part of the same outbreak have been compared to a reference genome. In the

* Correspondence: dave@cbs.dtu.dk
[2]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kgs Lyngby, Denmark
Full list of author information is available at the end of the article

future, it is likely that WGS will become a routine tool for identification and characterization of bacterial isolates, as hinted at in the first 'real-time' sequencing of the *E. coli* O104 outbreak in Germany in the summer of 2011 [10] and the *Vibrio cholerae* outbreak in Haiti in October 2010 [11]. This requires standard procedures for identifying variation and for analyzing similarities and differences.

Conserved genes are present across bacterial genomes of the same species (or genus). A fraction of these genes–those conserved in all (or most) of the genomes of a given bacterial taxonomic group–is called the 'core-genome' of that group. The core-genome can be identified either within a genus or species [3] and can be used to identify the variable genes in a given genome [12]. In addition, the conserved genes in general appear to evolve more slowly, and can be used for determining relationships among bacterial isolates [13].

Currently there are more than a hundred bacterial species for which sufficient genomic data are available to estimate the species core-genome (that is, there are at least three genomes sequenced from the same species) [14]. Among these, *Salmonella enterica* is a good candidate species for conserved gene identification because the genomes are quite similar [15]. Moreover, *S. enterica* is one of the most important food-borne pathogens and is responsible for global outbreaks [16] which makes international standard typing procedures of major importance in order to allow for global comparisons [17]. The *Salmonella* genus has only two species with sequenced genomes: *Salmonella bongori* and *Salmonella enterica*. In turn, *S. enterica* is divided into 6 sub-species: *enterica, salamae, arizonae, diarizone, houtenae* and *indica*. Presently, *S. enterica* is classified into more than 2,500 serotypes [18].

In order to investigate an outbreak caused by *Salmonella*, characterization of *Salmonella* isolates from genome data is a crucial step. *Salmonella* genomes are highly similar, particularly within subspecies *enterica*, where little variance exists in the genomes [15]. This high similarity presents a challenge for typing and classification.

In their pioneering work Tettelin *et al.* [1] defined the core genes of a species by being those genes found present in (nearly) all known members of the species. Since then others have studied core and pan genomes at the genus level or even at the kingdom level [19], but for our purposes the original definition at the species level is suitable. In this work we identify the core genes within *S. enterica* genomes and determine variation between the different available genomes, both in terms of sequence and presence/absence of non-core genes; in the latter case using a method originally published by Snipen & Ussery [20]. We evaluate the value of different approaches for classification of isolates in epidemiological settings and compare our findings to currently used sequencing methods, both in long term trend analysis and outbreak investigations.

Results and discussion

The 73 *Salmonella* genomes used in this study are summarized in Additional file 1: Table S1. The set comprises 21 completed genomes and 52 nearly completed genomes. Of these, 35 genomes are closely-related *S.* Montevideo strains pertaining to an outbreak of salmonellosis from Italian-style spiced meat [21]. All genomes were retrived from GenBank [22] except *S.* Typhimurium str. DT104, which was received from the Sanger Institute's bacterial genome database. All *Salmonella* genomes are from subspecies enterica with the exception of the single *S. enterica subsp. Arizonae*.

Evaluation of traditional bacterial sequence-based typing

The ribosomal genes are essential for the survival of all cells, and their structure cannot change much because of their involvement in protein synthesis [23]. Thus, 16S rRNA genes are highly conserved among isolates belonging to the same bacterial species [4]. Exceptions may be *N. meningitidis* [24] and *Mycoplasma* [25]. However, due to limited variation within a given species, the 16S sequencing is often not useful for epidemiological studies, where the classification of highly similar strains is needed. Jacobsen *et al.* shows a phylogenetic tree based on 16S rRNA genes, extracted from 26 *Salmonella enterica* genomes, using RNAmmer [15,26]. As expected, there is not sufficient resolution to distinguish among the *Salmonella* subspecies *enterica*.

Genes such as *rpoB* or *sodA* have been suggested as substitutes for 16S rRNA and have shown improved efficacy in species identification [27], although it remains unlikely that a single gene can always reflect the subtle differences between genomes of the same species.

The limitations of using a single gene may be improved by the simultaneous analysis of multiple genes. Multi Locus Sequence Typing (MLST) has found wide applications, especially in phylogenetic studies and is most commonly based on seven housekeeping genes - each bacterial species having its own set. For *Salmonella* these are: *aroC, dnaN, hemD, hisD, purE, sucA* and *thrA* http://www.mlst.net. A MLST tree, based on an *in silico* analysis of the 73 available *Salmonella enterica* genomes in Genbank, is shown in Figure 1. Strains of the same serovar generally cluster into distinct groups, although exceptions exist; for example the *S.* Weltevreden str. HI_N05-537 is mixed with *S.* Montivideo. Futhermore, recent work on 61 sequenced *E. coli* genomes [4], found that the 16S rRNA tree cannot resolve well within the genus level and also that MLST cannot differentiate pathogenic strains from non- pathogenic strains. Still, MLST has proven useful for long-term analysis of population structures, but often fails

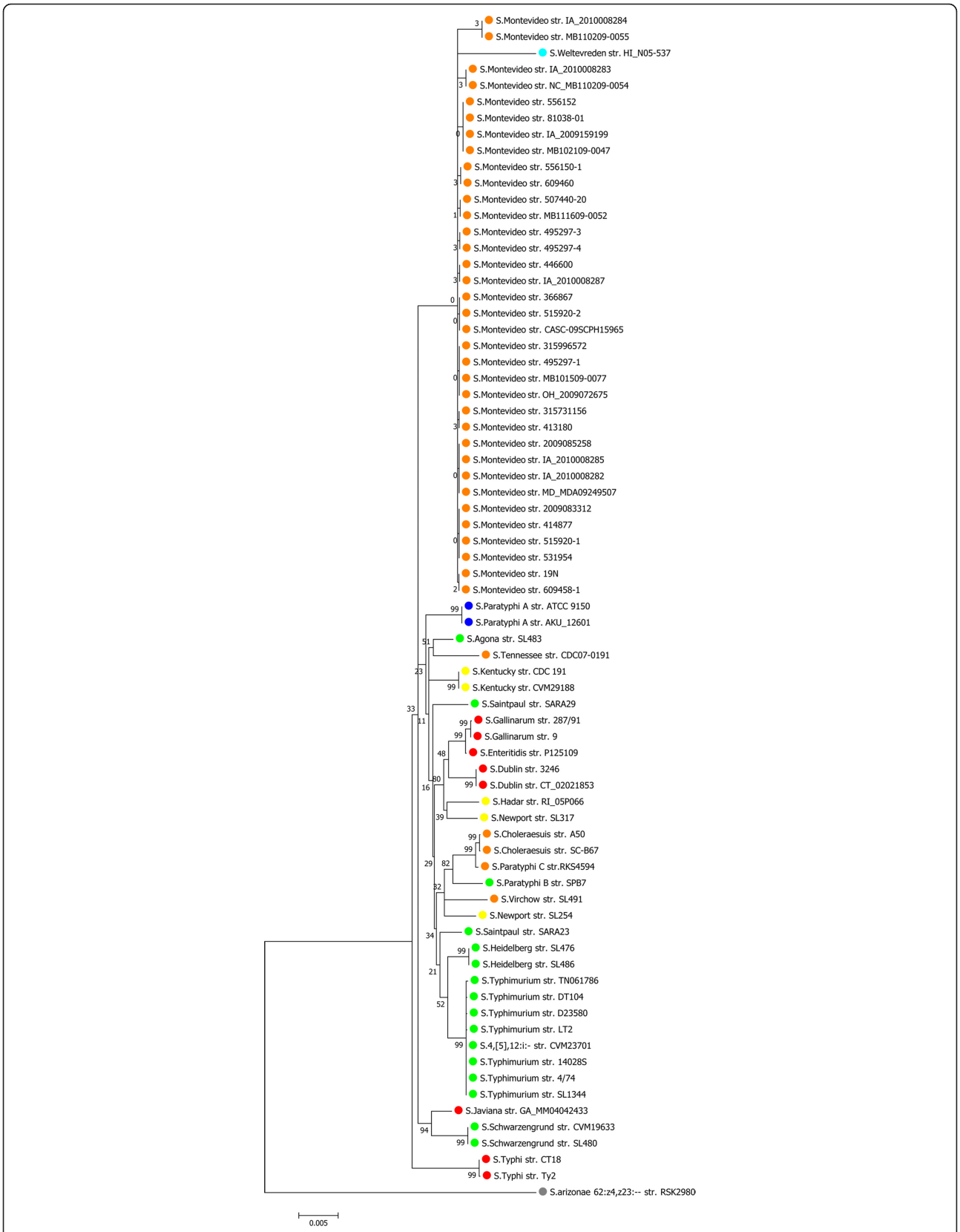


Figure 1 *In silico* MLST tree. Seven housekeeping genes were extracted from *Salmonella* genomes. Concatenated sequences were aligned by MUSCLE. The phylogenetic trees were generated by MEGA5 using bootstrap maximum likelihood method. Each color represents a different serogroup (O antigen). The confidence value is the bootstrap value calculated by sampling with replacement from the multiple sequence alignment.

to detect differences between closely related strains [28]. Indeed, improved MLST schemes that include more than 7 genes have been suggested [4].

For *Salmonella*, sequencing specific short repeats and virulence genes have recently been suggested as an alternative and improved method for typing of *S. Enteritidis* [29]. The usefulness of this approach in epidemiological studies and typing is currently unknown, although the choice of repeats must be tailored for the specific bacterial species studies.

Identification of core genes

Determining gene conservation across multiple genomes is not overly difficult, but certain choices must be made which will affect the final outcome. Using a previously published method [20,30,31] which employs single-linkage clustering on top of BLASTp alignments, sets of pan- and core-genomes were estimated, based on all 73 *Salmonella* genomes. The progression of the pan- and core-genomes is shown in Figure 2A. The number of novel gene clusters in the pan-genome gradually increases when more genomes are considered, while the number of conserved gene clusters constituting the core genome decreases slightly. When all *Salmonella* genomes have been considered, there are 10,581 pan gene clusters and 2,882 core gene clusters (Additional file 2) in species *enterica*. In the step going from *S*. Typhimurium to *S*. Typhi, the number of core genes drops suddenly, most likely because the *S*. Typhi genome has undergone considerable pseudogene formation resulting in gene loss [32]. The number of core genes drops again when adding a genome of the subspecies *arizonae* which is associated with cold-blooded animals. This technique has previously been applied successfully in finding core genomes for Proteobacteria genera *Burkholderia* [33], *Escherichia coli* [4], *Vibrionaceae* [34] and *Campylobacter jenuni* [30], as well as Bacteroides [35] and Lactic acid bacteria [36].

Genomic variation within the core genes

The core genes as calculated above were used for constructing a gene variation plot by performing all-against-all BLAST alignments between 2,882 core gene clusters and all 73 *Salmonella enterica* genomes. The resulting average identities within each core gene cluster is displayed in Figure 2B. From this figure, the average percent identity was very high (> 98%) in most of the core genes, but dropped sharply for around 5% of the core genes. From this plot, the identified core genes can be divided into two categories: a small group of highly variable genes and the majority of genes which show little variation.

For the highly variable core genes, the variation in amino acid sequences (Figure 2B, green dots) was higher than for the nucleotide sequences (Figure 2B, red dots), whereas the opposite was the case for the more conserved core genes. This indicates that for core genes with low variation there is a selection against mutations leading to amino acid changes, whereas for the highly variable genes, positive selection for amino acid changes seems to be the case. In order to confirm these hypothesis, the approximation of dN/dS has been performed by dividing the number of non-synonymous changes per non-synonymous sites with the number of synnonymous changes per synonymous sites [37] using *S*. Typhimurium str. LT2 as a reference genome. The median dN/dS ratio for conserved and highly variable core genes are 1.0 and 1.25 respectively. Therefore, the amino acid changes in highly variable core genes might be due to an increase in positive selection at some sites. Nonetheless, the importance of this needs to be confirmed by additional analysis, although one could imagine, for example, a selective pressure to vary the surface proteins to avoid immune response.

The seven genes used for MLST are marked in the Figure 2B, and are scattered throughout the highly conserved part of the core genes (Figure 2B, black dots) and, as expected, little variation exists in these genes. Including core genes from both the highly conserved and variable regions might be beneficial in evolution studies. On the one hand, the more slowly evolving genes are useful in distinguishing between divergent and convergent evolution, while faster evolving genes can help in strain identification.

Functional analysis of conserved genes

In order to determine the functional profile of core genes, the core gene clusters were aligned against UniProt [30]. Functional profiles were determined based on Gene Ontology (GO) terms and visualized in Figure 3. Though the difference is generally small, some terms common in conserved core genes tend to be less frequent in highly variable core genes; for example, electron carrier activity, structural molecule activity and metallochaperone activity. These functions are essential for living cells and are therefore enriched in conserved core genes. On the other hand, highly variable core genes encode many proteins that are associated with the extracellular region. In general, genes located outside the cell are known to be more variable [38].

Consensus tree based on core gene clusters

Figure 4 shows a phylogenetic tree generated from the sequence of all 2,882 *Salmonella* core gene clusters. The tree generally divides the serotypes up well, but the bootstrap value in several branches is very low. This uncertainty could be due to the large number of core gene trees being analyzed individually; the low bootstrap values near the root reflect a lack of consensus at the higher levels. In contrast, the low bootstrap values found in *S*. Montevideo strains likely reflect uncertainty due to the high similarity of gene sequence of the clonal

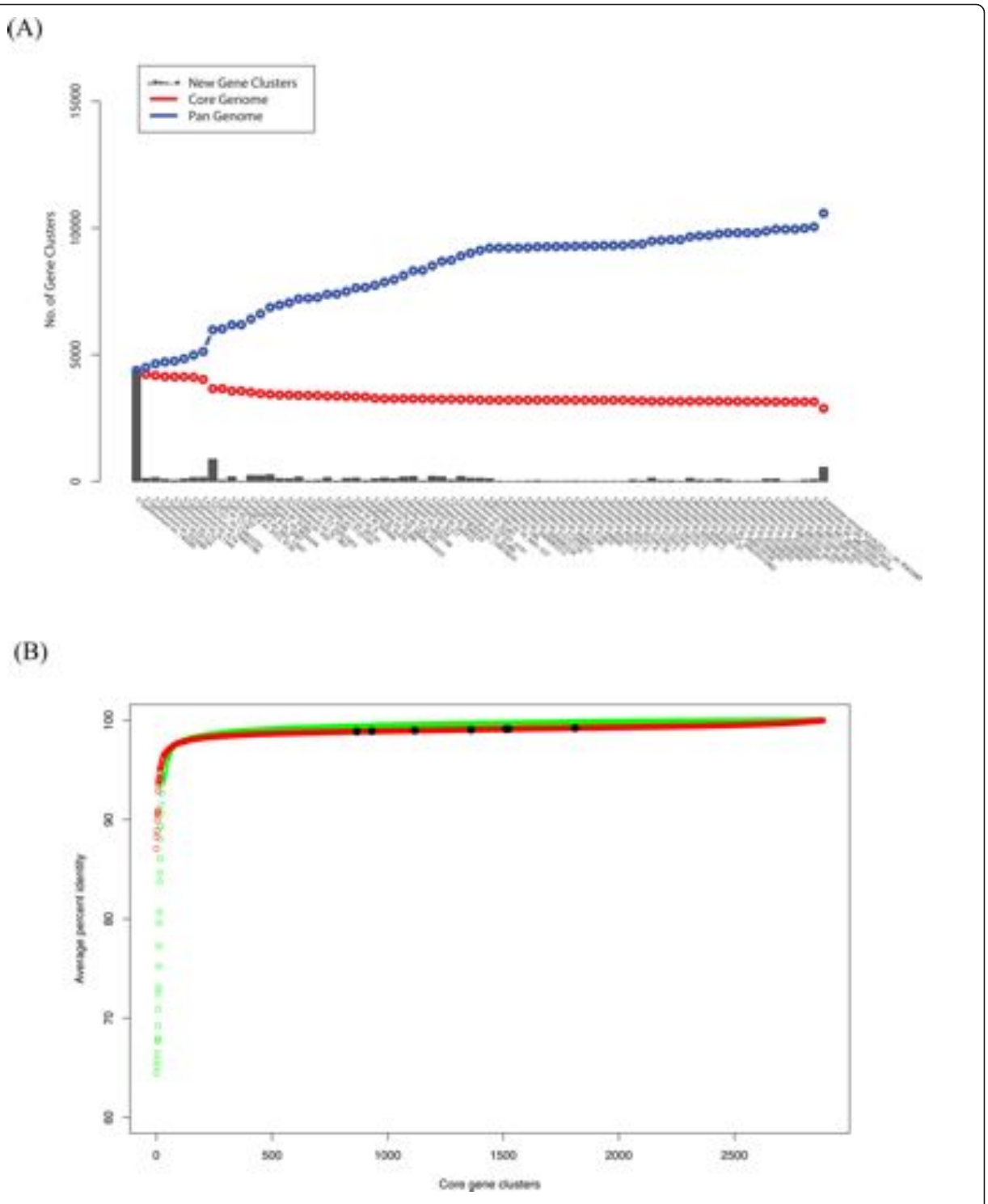


Figure 2 Pan- core-genome plot and variation plot. (A) Pan- and core-genome plot of 73 *Salmonella enterica*. The plot shows an increase of the pan-genome (blue line) and a decrease of the core-genome (red line) as more genomes are added. The last points show the total number of gene clusters in the pan-genome and the core-genome. **(B)** Variation plot. This plot shows the variation within core gene clusters in amino acid levels (green dots) and nucleotide levels (red dots). Black dots show the distribution of housekeeping genes in the core genes. The Y- and X-axes represent average percent identity and numerical core gene cluster name respectively.

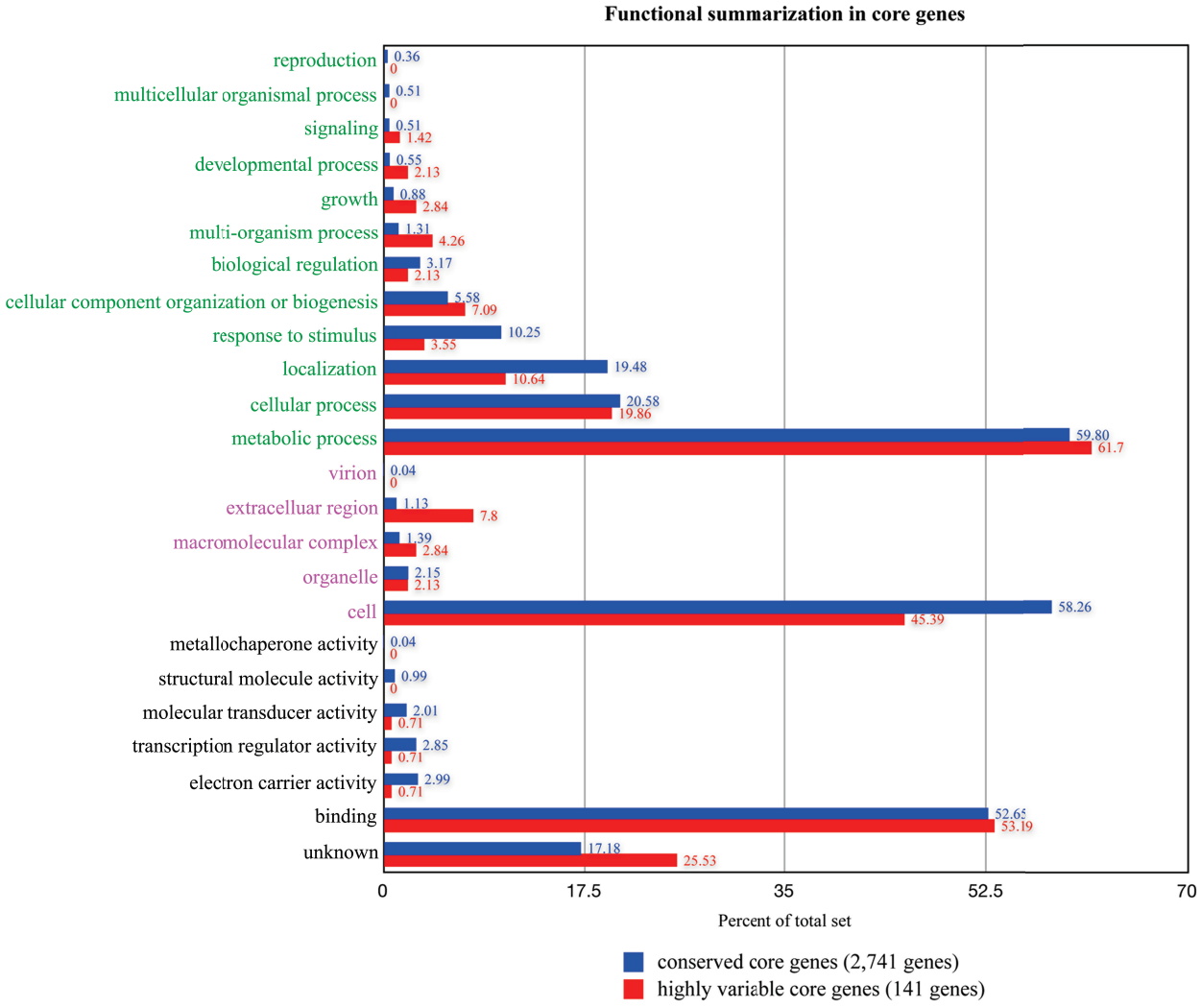


Figure 3 Gene Ontology term summary of core genes. Gene Ontology terms for conserved core genes (blue bars) and highly variable core genes (red bars) are shown in 3 categories (from top to bottom): biological processes (green labels), cellular component (pink labels) and molecular function (black labels). GO are assigned from blast all-against-all between core genes and protein sequences from Uniprot based on 50/50 rule. All conclusions drawn about the variable set are relative to the fraction of like sequences in the conserved set, and not in any way absolute.

outbreak. All *S.* Montevideo strains sequenced were from a single outbreak [21] and as expected this analysis confirmed the almost complete identity of these isolates.

A previous study described that there are 69 genes unique to *Salmonella* [39]. Instead of using all core genes, we generated a consensus tree based on these 69 *Salmonella*-specific genes (Additional file 3: Figure S1). We also constructed an additional four consensus trees based on sets of 69 core genes randomly picked from different areas in the variation plot (Figure 2B): from a mixture of high, medium and low variable core genes (Additional file 4: Figure S2), from medium variable core genes (Additional file 5: Figure S3), from highly variable core genes (Additional file 6: Figure S4) and from the area where the curve decreases in the variation plot (Additional file 7: Figure S5). The appearance of these 5 consensus trees was similar to the tree from Figure 4, with two exceptions: the trees based on the 69 specific genes (Additional file 3: Figure S1) and the highly variable core genes (Additional file 6: Figure S4). In the former, *S. arizonae*, which is not part of the subspecies *enterica*, was still mixed in with other *enterica*, while for the latter, *S.* Agona str. SL483 clustered away from the other subspecies *enterica*. Thus, based on these results, it appears that using only *Salmonella* unique genes or highly variable genes does not provide phylogenetically useful information and should probably not be used for future WGS studies. Comparisons using more genomes in more species can further test this.

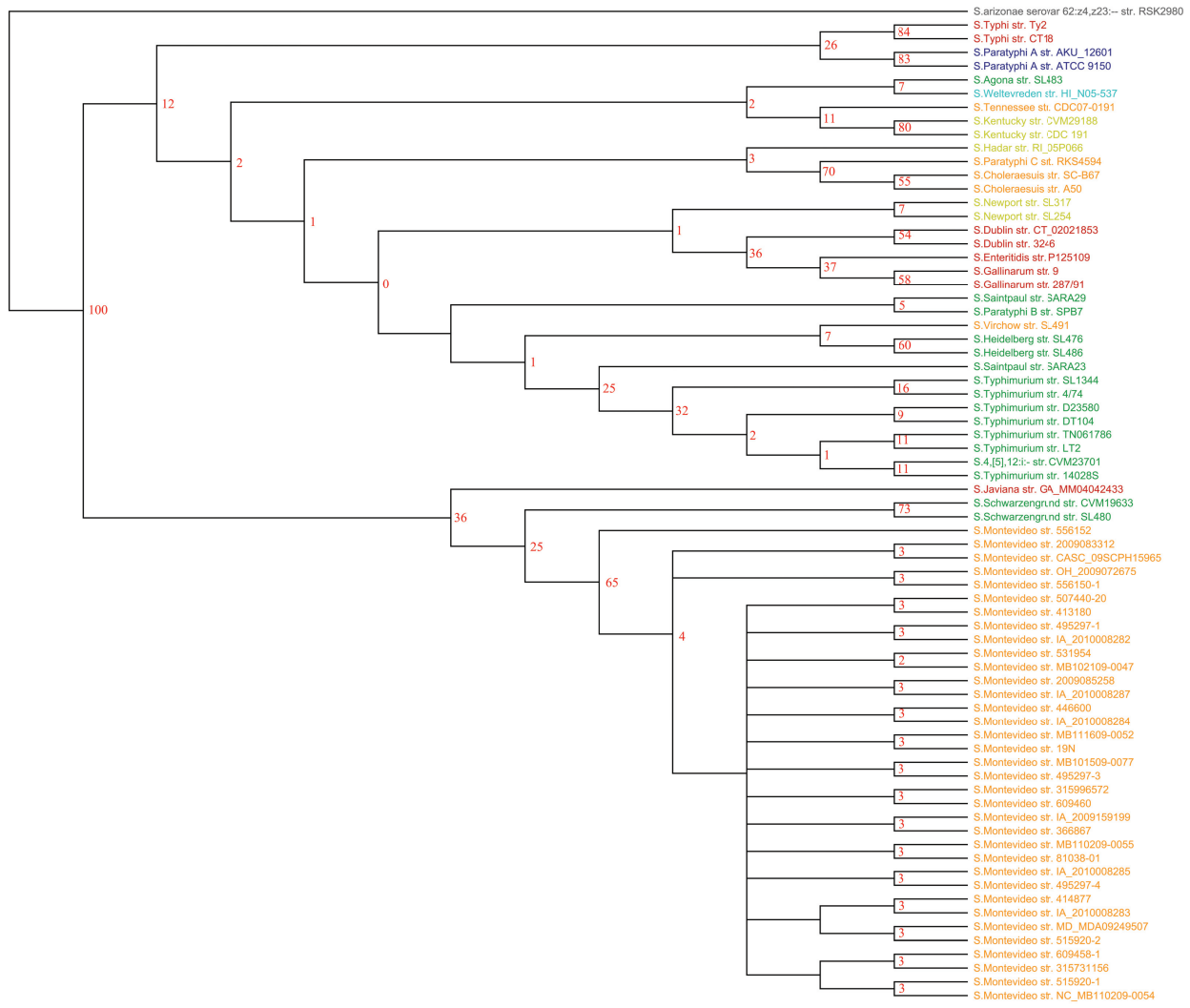


Figure 4 Consensus tree based on 2,882 core gene clusters. Phylogenetic trees were constructed from all core genes using PAUP. All trees were combined and the consensus trees were generated using the Phylip software package. The percentage of branches present in all trees is shown. The colors represent different serogroups, as in Figure 1.

Pan-genome tree

In principle, genome similarity is not only measurable by shared genes, but also by the absence of genes. Figure 5 is another tree, based on gene presence/absence across all the *Salmonella* genomes [20]. This tree bears a striking resemblance to the consensus tree based on core genes (Figure 4), although the bootstrap values are higher in many of the branches, especially near the root. Of all methods investigated in this study, the pan-genome tree presents itself as the best solution for a tree that can resolve strain differences in a biologically meaningful way, even if it would be expected to correlate more with phenotype than phylogeny. It is, however, important to note that creating pan-genome trees requires higher quality sequencing data and assemblies than what are typically obtained using short reads from second-generation sequencing methodologies. Even so, we have found that pan-genome trees with good correspondence to known bacterial types can be constructed from Solexa data (100 bp reads), if care is taken to ensure good assembly and gene finding (data not shown).

The power to discriminate between variants differs between the methods used. The phylogenetic analysis for the MLST tree is based on the identified informative sites among the seven housekeeping genes, for the pan-genome tree on presence and absence of genes and for the consensus tree based on the informative sites of core gene clusters from alignments of all core genes trees. The number of infomative sites for *in silico* MLST tree, pan-genome tree and consensus tree based on core gene clusters were

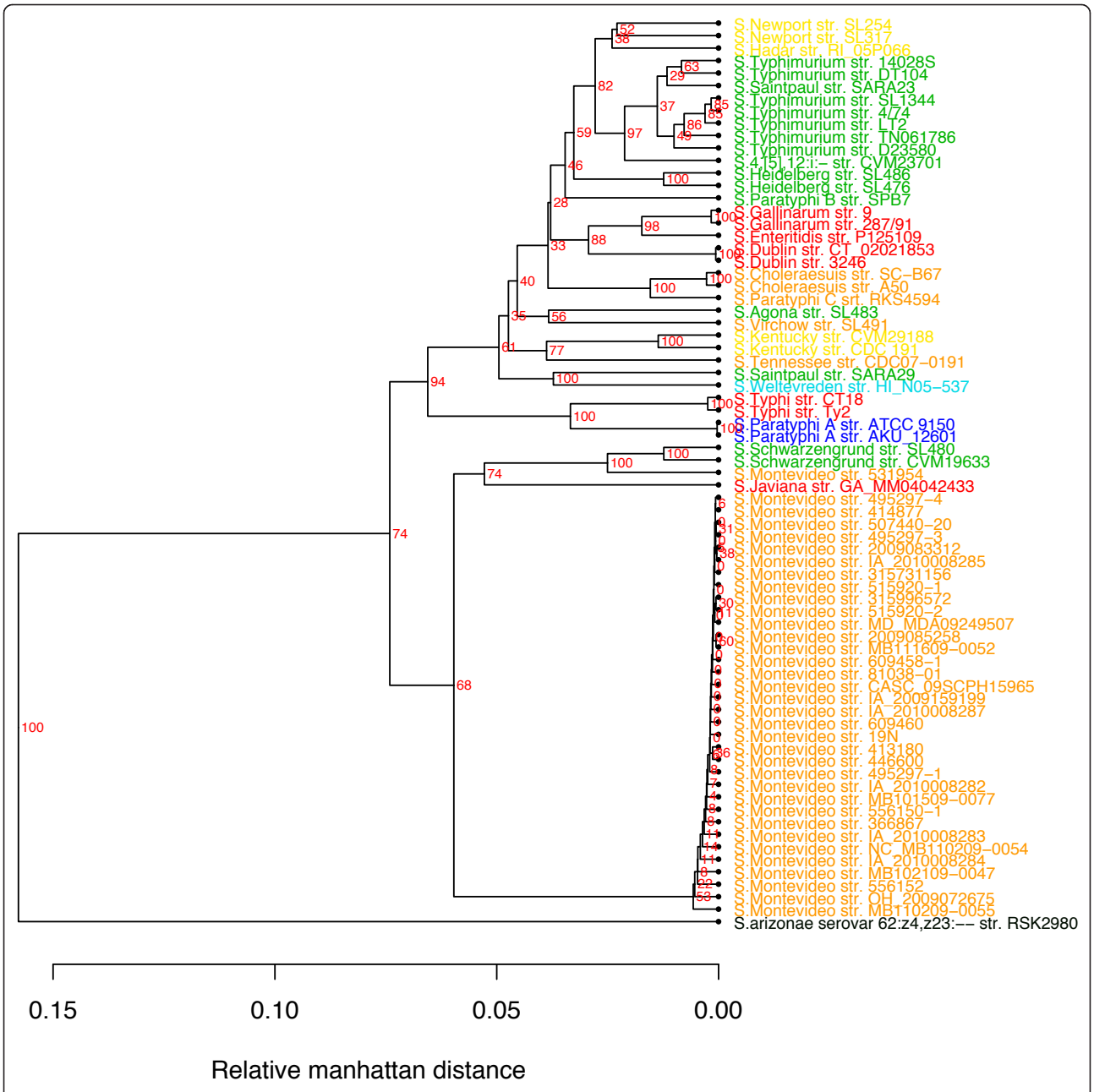


Figure 5 Pan-genome tree. This tree does not produce a sequence-based alignment tree but it is generated from the presence or absence of gene clusters across all *Salmonella* genomes [31]. The bootstrap values are shown in red.

877 bp (10,008 total base-pairs in the seven genes), 7,699 genes (10,581 total genes) and 880,832 bp (2,868,821 bp in all core genes), respectively. The pan genome and core gene analysis were based on much more variation than the MLST analysis and have a much stronger power to discriminate closely related strains.

Conclusions

Bacterial typing should provide meaningful information for both epidemiological and evolutionary studies. For epidemiology, the ability to differentiate unrelated isolates (discriminatory power) and the ability to cluster related isolates are crucial. 16S rRNA and the MLST genes rarely provide separation between closely related strains. The performance of the pan-genome tree, however, is valid for epidemiological investigation in both discriminatory and clustering abilities. One caveat is that this method depends on good quality genomic data.

Comparative genomics can determine the conserved genes (core-genome) among bacterial genomes at either

genus or species level. Genomic variation within the core-genome can then be used to reveal highly variable genes (fast evolving genes) and conserved genes (slow evolving genes). These core genes are useful for investigating molecular evolution and remain useful as candidate genes for bacterial genome typing–even if they cannot be expected to differentiate highly similar isolates from e.g. outbreak cases, such is not always desirable. Even in cases where a deeper distinction of isolates is of interest, e.g. in mapping outbreaks, core genes might still be useful as a reference fragment for SNPs calling instead of using whole genome analysis. However, in term of computational costs, the consensus tree based on core genes requires more computational time than the other methods.

In the near future, global real-time surveillance of *Salmonella* and other pathogens giving simultaneous information on population structure and evolution, as well as outbreak detection, may well be possible.

Methods

Salmonella genome data and gene annotation

From public genome databases (NCBI and Sanger Institute's bacterial genome databases), 83 *Salmonella enterica* genomes available at the time (April, 2011) were downloaded. These genomes consisted of 21 completed genomes and 62 draft genomes. Due to the large number of contigs in some genomes, only 73 genomes were selected for this study (Additional file 1: Table 1). The gene finder Prodigal was used on DNA sequences of all genomes to eliminate biases in annotation quality and to standardize the genes found in all genomes [15]. Gene clusters were then inferred according to [15,20,30]

In silico MLST trees

The *in silico* MLST tree was constructed from seven housekeeping genes: *aroC, dnaN, hemD, hisD, purE, sucA* and *thrA* http://www.mlst.net. These genes were extracted from *Salmonella* genomes and concatenated. The concatenated sequences were aligned using MUSCLE [40]. Phylogenetic trees were generated by MEGA5 using the maximum likelihood method [41]. The confidence value is, in this case, the same as the bootstrap value, calculated by sampling with replacement from the multiple sequence alignments [42]. Thus, the *in silico* MLST differs from traditional MLST in that complete genes are used and not just the MLST alleles. However, since the alleles typically cover the majority of the genes, the difference is small.

Consensus trees

All core gene clusters from 73 *Salmonella* genomes were used for generating a consensus tree. Multiple alignments for each core gene cluster from all strains were performed using MUSCLE [40]. A phylogenetic tree for each core gene was generated using PAUP [43]. The Phylip package was used to construct the consensus tree from all the trees [44]. The bootstrap values are shown in the consensus tree.

GO annotation

The core gene clusters were compared in an all-against-all BLAST with protein sequences from UniProt based on the '50/50 rule' [30]. Functional profiles were summarized from BLAST results by mapping UniProt IDs to Gene Ontology (GO) terms. Mapping GO parental terms were performed using publicly available GO-PERL modules for searching through a graph structure of ontology data [45,46]

Pan-genome trees

The Pan-genome matrix consists of gene clusters (rows) and genomes (columns). The absence and presence of genes across genomes are represented by 0's and 1's respectively. The relative Manhattan distance between genomes was calculated and used for hierarchical clustering. The bootstrap values are calculated in order to represent the confidence of branches [20].

Additional material

Additional file 1: Table S1 List of *Salmonella* genomes used in this study.

Additional file 2: Core gene clusters. This file contains 2,882 *Salmonella* core genes in FASTA format.

Additional file 3: Figure S1 Consensus tree based on 69 specific *Salmonella* genes.

Additional file 4: Figure S2 Consensus tree based on 69 *Salmonella* core genes randomly picked up from high, medium and low variable core genes.

Additional file 5: Figure S3 Consensus tree based on 69 Salmonella core genes randomly picked up from medium variable core genes.

Additional file 6: Figure S4 Consensus tree based on 69 Salmonella core genes randomly picked up from highly variable core genes.

Additional file 7: Figure S5 Consensus tree based on 69 Salmonella core genes randomly picked up from decreasing curve in the variation plot.

Acknowledgements

This study was supported by the Center for Genomic Epidemiology (09-067103/DSF) http://www.genomicepidemiology.org and by grant 3304-FVFP-08- from the Danish Food Industry Agency. PL and OL would like to acknowledge funding from the Technical University of Denmark. The authors would like to thank Colleen Ussery for editorial assistance in preparing the manuscript.

Author details

[1]National Food Institute, Building 204, The Technical University of Denmark, 2800 Kgs Lyngby, Denmark. [2]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kgs Lyngby, Denmark.

Authors' contributions

PL planned the study, carried out all bioinformatics analysis and drafted the manuscript. OL participated in consensus tree based on core genes. CF participated in the planning of the study, the core genes identification and drafted the manuscript. FMA supervised and planned the study and drafted the manuscript. DWU supported the supervision, participated in the design of the study and drafted the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 30 September 2011 Accepted: 12 March 2012
Published: 12 March 2012

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit Y, Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactia*: Implications for the microbial "pan-genome".** *Proc Natl Acad Sci USA* 2005, **102(39)**:13950-13955.
2. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussery DW: **Ten years of bacterial genome sequencing: comparative-genomics- based discoveries.** *Funct Integr Genomics* 2006, **6**:165-185.
3. Malorny B: **New Approaches in Subspecies-level *Salmonella* Classification.** In *Salmonella From Genome to Function.* Edited by: Porwollik S. Norwich United Kingdom: Caister Academic Press; 2011:1-23.
4. Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 Sequenced Escherichia coli Genomes.** *Microb Ecol* 2010, **60(4)**:708-720.
5. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA During Hospital Transmission and Intercontinental Spread.** *Science* 2010, **327(5964)**:469-474.
6. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15(6)**:L589-L594.
7. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *N Engl J Med* 2011, **364(8)**:730-739.
8. Rasko DA, Worsham PL, Abshire TG, Stanley ST, Bannan JD, Wilson MR, Langham RJ, Decker RS, Jiang L, Read TD, Phillippy AM, Salzberg SL, Pop M, Van Ert MN, Kenefic LJ, Keim PS, Fraser-Liggett CM, Ravel J: ***Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation.** *Proc Natl Acad Sci* 2011, **108(12)**:5027-5030.
9. Pallen MJ, Loman NJ, Penn CW: **High-throughput sequencing and clinical microbiology: progress, opportunities and challenges.** *Curr Opin Microbiol* 2010, **13(5)**:625-631.
10. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H: **Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia col* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology.** *PLoS One* 2011, **6(7)**:e22751.
11. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK: **The origin of the Haitian cholera outbreak strain.** *N Engl J Med* 2011, **364(1)**:33-42.
12. Adékambi T, Butler RW, Hanrahan F, Delcher AL, Drancourt M, Shinnick TM: **Core gene set as the basis of multilocus sequence analysis of the subclass Actinobacteridae.** *PLoS One* 2011, **6(3)**:e14792.
13. Urwin R, Maiden MC: **Multi-locus sequence typing: a tool for global epidemiology.** *Trends Microbiol* 2003, **11(10)**:479-487.
14. Kyrpides NC: **Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream.** *Nat Biotechnol* 2009, **27(7)**:627-632.
15. Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C: **The *Salmonella enteric* Pan-genome.** *Microb Ecol* 2011, **62(3)**:487-504.
16. Foley SL, Zhao S, Walker RD: **Comparison of molecular typing methods for the differentiation of salmonella foodborne pathogens.** *Foodborne Pathog Dis* 2007, **4(3)**:253-276.
17. Boxrud D, Monson T, Stiles T, Besser J: **The role, challenges, and support of pulsenet laboratories in detecting foodborne disease outbreaks.** *Public Health Rep* 2010, **125(Suppl 2)**:57-62.
18. Popoff MY, Le Minor L: **Taxonomy of the genus *Salmonella*. Changes in serovars nomenclature.** In *Antigenic formulas of the Salmonella serovars, 7th revision.* Edited by: Popoff MY, Le Minor L. Paris, France: WHO Collaborating Centre for Reference and Research on Salmonella. Institut Pasteur; 1997:5.
19. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genomes.** *Trends Genet* 2009, **25(3)**:107-110.
20. Snipen L, Ussery DW: **Standard operation procedure for computing pangenome trees.** *Stand Genomics Sci* 2009, **2**:135-141.
21. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R: **Identification of a Salmonellosis outbreak by means of molecular sequencing.** *N Engl J Med* 2011, **364(10)**:981-982.
22. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2011, **39**:D32-D37.
23. Woose CR: **Bacterial evolution.** *Microbiol Rev* 1987, **51(2)**:221-271.
24. Sacchi CT, Whitney AM, Reeves MW, Mayer LW, Popovic T: **Sequence diversity of *Neisseria meningitidi* 16S rRNA genes and use of 16S rRNA gene sequencing as a molecular subtyping tool.** *J Clin Microbiol* 2002, **40(12)**:4520-4527.
25. Königsson MH, Bölske G, Johansson KE: **Intraspecific variation in the 16S rRNA gene sequences of *Mycoplasma agalactia* and *Mycoplasma bovi* strains.** *Vet Microbiol* 2002, **85(3)**:209-220.
26. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35(9)**:3100-3108.
27. De Clerck E, De Vos P: **Genotypic diversity among *Bacillus licheniformi* strains from various sources.** *FEMS Microbiol Lett* 2004, **231(1)**:91-98.
28. Li W, Raoult D, Fournier PE: **Bacterial strain typing in the genomic era.** *FEMS Microbiol Rev* 2009, **33(5)**:892-916.
29. Liu F, Kariyawasam S, Jayarao BM, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG: **Subtyping Salmonella enterica Serovar Enteritidis Isolates from Different Sources by Using Sequence Typing Based on Virulence Genes and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs).** *Appl Environ Microbiol* 2011, **77(13)**:4520-4526.
30. Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, Hallin PF, Newell DG, Toszeghy M, Ridley A, Manning G, Ussery DW: **Genomic characterization of *Camphylobacter jejun* M1.** *PLoS One* 2010, **5(8)**:e12253.
31. Ussery DW, Wassenaar TM, Borini S: *Computing for Comparative Genomics: Bioinformatics for Microbiologists (Computational Series)* London: Springer Verlag; 2008.
32. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, Bhutta ZA, Quail MA, Norbertczak H, Walker D, Simmonds M, White B, Bason N, Mungall K, Dougan G, Parkhill J: **Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi.** *BMC Genomics* 2009, **10**:36.
33. Ussery DW, Kill K, Lagesen K, Sicheritz-Ponten T, Bohlin J, Wassenaar TM: **The Genus Burkholderia: Analysis of 56 Genomic Sequences. Microbial Pathogenomics.** In *Microbial Pathogenomics.* Edited by: Reuse Hd, Bereswill S. Basel, Karger; 2009:140-157.
34. Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, Ussery DW: **On the Origins of a Vibrio Species.** *Microb Ecol* 2010, **59(1)**:1-13.
35. Karlsson FH, Ussery DW, Nielsen J, Nookaew I: **A closer look at bacteroides: phylogenetic relationship and genomic implications of a life in the human gut.** *Microb Ecol* 2011, **61(3)**:473-485.
36. Lukjancenko O, Ussery DW, Wassenaar TM: **Comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera.** *Microb Ecol* 2011.
37. Yi S: **Synonymous and Nonsynonymous Rates.** *eLS* 2007, doi: 10.1002/9780470015902.a0005110.pub2.
38. Julenius K, Pedersen AG: **Protein evolution is faster outside the cell.** *Mol Biol Evol* 2006, **23(11)**:2039-2048.

39. Lukjancenko O, Ussery DW: **Design of an Enterobacteriaceae Pan-Genome Microarray Chip.** *Proceeding of CSBio 2010: Thailand* 2010, **115**:174-189.
40. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
41. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28(10)**:2731-2739.
42. Wróbel B: **Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods.** *J Appl Genet* 2008, **49(1)**:49-67.
43. Swofford DL: *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4* Sunderland: Sinauer Associates; 2004.
44. Felsenstein J: *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences* Seattle: University of Washington; 2005.
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
46. Leekitcharoenphon P, Taweemuang U, Palittapongarnpim P, Kotewong R, Supasiri T, Sonthayanon B: **Predicted sub-populations in a marine shrimp proteome as revealed by combined EST and cDNA data from multiple *Penaeu* species.** *BMC Res Notes* 2010, **3**:295.

doi:10.1186/1471-2164-13-88
Cite this article as: Leekitcharoenphon *et al.*: Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* 2012 **13**:88.

Article II

Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*.

ARTICLE II

Pimlapas Leekitcharoenphon,[1,2*] Eva M. Nielsen,[3] Rolf S. Kaas,[1,2] Ole Lund,[2] Frank M. Aarestrup,[1]

[1] Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs Lyngby, Denmark

[3] Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark

*Corresponding author: **Pimlapas Leekitcharoenphon,**
Division for Epidemiology and Microbial Genomics,
National Food Institute, Technical University of Denmark,
Kgs. Lyngby, Denmark
E-mail: pile@food.dtu.dk

PLoS One 2014 Feb 4;9(2):e87991

OPEN ACCESS Freely available online

PLOS ONE

Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*

Pimlapas Leekitcharoenphon[1,2]*, Eva M. Nielsen[3], Rolf S. Kaas[1,2], Ole Lund[2], Frank M. Aarestrup[1]

1 Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark, **2** Department of System Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark, **3** Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark

Abstract

Salmonella enterica is a common cause of minor and large food borne outbreaks. To achieve successful and nearly 'real-time' monitoring and identification of outbreaks, reliable sub-typing is essential. Whole genome sequencing (WGS) shows great promises for using as a routine epidemiological typing tool. Here we evaluate WGS for typing of *S.* Typhimurium including different approaches for analyzing and comparing the data. A collection of 34 *S.* Typhimurium isolates was sequenced. This consisted of 18 isolates from six outbreaks and 16 epidemiologically unrelated background strains. In addition, 8 *S.* Enteritidis and 5 *S.* Derby were also sequenced and used for comparison. A number of different bioinformatics approaches were applied on the data; including pan-genome tree, k-mer tree, nucleotide difference tree and SNP tree. The outcome of each approach was evaluated in relation to the association of the isolates to specific outbreaks. The pan-genome tree clustered 65% of the *S.* Typhimurium isolates according to the pre-defined epidemiology, the k-mer tree 88%, the nucleotide difference tree 100% and the SNP tree 100% of the strains within *S.* Typhimurium. The resulting outcome of the four phylogenetic analyses were also compared to PFGE reveling that WGS typing achieved the greater performance than the traditional method. In conclusion, for *S.* Typhimurium, SNP analysis and nucleotide difference approach of WGS data seem to be the superior methods for epidemiological typing compared to other phylogenetic analytic approaches that may be used on WGS. These approaches were also superior to the more classical typing method, PFGE. Our study also indicates that WGS alone is insufficient to determine whether strains are related or un-related to outbreaks. This still requires the combination of epidemiological data and whole genome sequencing results.

Citation: Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM (2014) Evaluation of Whole Genome Sequencing for Outbreak Detection of *Salmonella enterica*. PLoS ONE 9(2): e87991. doi:10.1371/journal.pone.0087991

Editor: Jose Alejandro Chabalgoity, Facultad de Medicina, Uruguay

Received October 21, 2013; **Accepted** January 2, 2014; **Published** February 4, 2014

Funding: This study was supported by the Center for Genomic Epidemiology (09- 067103/DSF) http://www.genomicepidemiology.org. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pile@food.dtu.dk

Introduction

Salmonella is a common cause of infectious disease in human and animals. *Salmonella* is classically divided into species *S.bongori* and *S.enterica*; the latter further divided into more than 2,500 different serotypes [1,2]. It is, however, only a limited number of serovars that are responsible for most infections and in Europe, the most prevalent *S.enterica* serovars isolated from humans are Enteritidis and Typhimurium, responsible for over 75% of the human cases of salmonellosis [3]. *Salmonella* infections can occur as minor and major foodborne outbreaks (major outbreak - an outbreak that attracts intensive publicity). In order to elucidate the epidemiology and implement the control programs, reliable and rapid sub-typing is essential [4,5]. Today, different typing methods are commonly used as a central part of the detection and investigation of *Salmonella* outbreaks, for instance, serotyping, phage typing, pulse-field gel electrophoresis (PFGE) and multilocus variable number of tandem repeat analysis (MLVA) [6–8]. PFGE has been the gold standard for epidemiological investigations of foodborne bacterial pathogens including *Salmonella* [9]. A drawback of PFGE is that it is unable to separate very closely related strains because the low rate of genetic variation does not significantly impact the electrophoretic mobility of a restriction fragment [6]. MLVA has major benefits in epidemiological surveillance of some *Salmonella* [10], but serotype specific protocols are needed for high discrimination.

During recent years the cost of whole genome sequencing (WGS) has decreased dramatically and the technology becomes increasingly available for routine use around the world [4,11]. Moreover, the speed of sequencing is decreasing from several days or weeks to perhaps hours for a bacterial genome in the near future [12]. The combination of low cost and high speed of WGS, opens an opportunity for WGS to become very useful and practical in various bacterial infectious studies [13–15] including the routine use in diagnostic and public health microbiology [12,16]. WGS has also been successfully used for elucidating the evolution of some *Salmonella* sub-types [15,17]. Nevertheless, prior to implementing WGS in routine surveillance, it is essential to evaluate it compared to traditional method and to determine which analytic approaches that might be most useful for a given bacterial species and sub-type.

This study was conducted to evaluate WGS for outbreak typing of *S.enterica*. A collection of presumed epidemiologically related and un-related *S.enterica* strains were sequenced and analyzed using four different bioinformatics approaches. The outcome was evaluated according to the pre-defined expected epidemiological

data and also compared to results obtained using the conventional typing method, PFGE.

Methods

Bacterial Isolates and Molecular Typing

Salmonella strains were derived from the Danish laboratory-based surveillance system of human gastrointestinal infections in 2000–2010. The procedures for isolation, identification, serotyping, antimicrobial susceptibility testing, PFGE and MLVA of the isolates included in this study have been described previously [9,18]. The *S.* Typhimurium collection consisted of 18 isolates from 6 previously described outbreaks or clusters, primarily defined by MLVA [9,10] and 16 strains that were expected to be epidemiologically un-related to the outbreaks. The outbreaks were selected to cover outbreaks that were restricted in time and location [10] as well as some epidemiologically challenging outbreaks (outbreak 1–3) that lasted several months [9]. The isolates from each outbreak/cluster were selected to include some of the known diversity within these (e.g. based on phage type, MLVA, PFGE as well as the time span of the outbreak). The 16 background strains were selected, so at least two isolates belonged to the same phage type as that of each of the 6 outbreaks. The set of *S.* Enteritidis consisted of 5 isolates from a couple of outbreaks and 3 background strains. The *S.* Derby collection comprised 3 isolates from a single outbreak and 2 background strains. Isolate information was included in Table 1.

Whole Genome Sequencing

The total set of forty-seven *Salmonella enterica* genomes was selected for multiplexed, paired-end sequencing on the Illumina GAIIx genome analyzer (Illumina, Inc., San Diego, CA). The procedures for DNA and library preparation including sequencing in this study have been described previously and according to Hendriksen *et al* [13]. The paired-end reads had read length at 101 bp. The genomic data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. ERP002633. The raw reads can be accessed online at http://www.ebi.ac.uk/ena/data/view/ERP002633. *De novo* short read assembly was performed on the set of raw reads using Velvet [19], which is a part of the pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [20,21]. The *de novo* assembly produced contigs with average N50 = 232,749.

A number of publicly available *Salmonella* genomic data were integrated to this study making total set of analyzed data rose to 271 genomes. A set of 39 *S.* Montevideo genomes was retrieved via Bioproject 61937 with the accession numbers AESR00000000-AESY00000000, AHIA00000000 and AHHT00000000 - AHHW00000000 [17]. Nine *S.* Heidelberg genomes were downloaded using the accession number AMBU00000000, AMBV00000000, AMBW00000000, AMBX00000000, AJGW00000000, AJGX00000000, AJGY00000000, AJGZ00000000, and AJHA00000000 [22,23]. A set of 71 *S.* Agona were received through EMBL genomic assemblies at www.ebi.ac.uk/ena (PRJEB1064-1135) [24]. A number of 105 *S.* Enteritidis genomes were retrieved via NCBI with the accession number AHUJ00000000- AHUR00000000, ALEA00000000-ALEZ00000000, ALFA00000000- ALFZ00000000, ALGA00000000-ALGZ00000000, ALHA00000000- ALHZ00000000 and ALIA00000000- ALID00000000 [25].

Pan-genome Tree

Pan-genome tree was constructed from the pan-genome matrix that composed of genes and genomes (*de novo* assembled genomes from this study) as rows and columns respectively. The matrix contains profile of 0′s and 1′s represented as the absence and presence of genes across genomes. The pan-genome tree was computed on the basis of distance between pan-genome profiles using a relative Manhattan distance. The tree can be formed by hierarchical clustering by employing an average linkage, corresponding to the Unweighted Pair-Group Method with Arithmetic mean (UPGMA) algorithm. The stability of the branching was illustrated via bootstrapping. This was implemented by re-sampling genes i.e. rows of the pan-matrix, and re-clustering these data. The bootstrap value for a split is the percentage of the re-sampled trees having a similar node, i.e. with the same two sets of leaves in the branches [26,27].

K-mer Tree

K-mer tree, alignment-free genome phylogeny, is constructed from the contiguous sequences of k bases called k-mers [28]. K can be any positive integer. In principle, sequences with high similarity likely share k-mers [29,30]. Based on this idea, the *de novo* assembled genomes were split into short sequences with the size of k (k-mers). If the k-mer size is tiny, the alignment specificity of k-mers will be low. If the k-mers are too large, they will be seldom aligned. K-mers were aligned against all the genomes. The number of hits or the frequency of k-mers across genomes was constructed as a matrix. The matrix consists of k-mers and genomes (rows and columns respectively) with the frequency of k-mers hits as a profile. The hierarchical clustering was performed in order to build the k-mer tree.

Nucleotide Difference Tree (ND Tree)

We used the well-studied *S.* Typhimurium str. LT2 as a reference genome (National Center for Biotechnology Information, accession: AE006468, length of 4,857,432 bp). The reference genome was split into k-mers of length 17 and stored in a hash table. Each read with a length of at least 50 was split into 17-mers overlapping by 16. K-mers from the read and its reverse complement were mapped until an ungapped alignment with a score of at least 50 was found using a match score of 1 and a mismatch score of −3.

When all reads had been mapped, the significance of the base call at each position was evaluated by calculating the number of reads X having the most common nucleotide at that position, and the number of reads Y supporting other nucleotides. A Z-score was calculated as $Z = (X-Y)/\mathrm{sqrt}(X+Y)$. The value of 1.96 was used as a threshold for Z corresponding to a p-value of 0.001. It was further required that $X>10*Y$.

Each pair of sequences was compared and the number of nucleotide differences in positions called in all sequences was counted. We obtained similar results by using a more strict threshold of $z = 3.29$, but then counting nucleotide differences at all positions called by both of the strains to be compared (data not shown). A matrix with these numbers was given as input to a UPGMA algorithm implemented in the neighbor program (http://evolution.genetics.washington.edu/phylip.html) in order to construct the tree. The ND tree approach was implemented as a pipeline tool on the Center for Genomic Epidemiology (http://www.cge.cbs.dtu.dk/services/NDtree/).

Identification of Core Genes

The set of 2,882 *Salmonella* core genes was downloaded from supplementary data of a previous publication [2]. This set of core genes (conserved genes) was estimated based on 73 publicly available *Salmonella* genomes using a previously published clustering method, which employs single-linkage clustering on top of

Table 1. Epidemiological information for the 47 *Salmonella* genomes used in this study (source: human).

ID	Serotype	Received date	Outbreak/ Background	Outbreak no.	Phage type	STTR9	STTR5	STTR6	STTR10	STTR3	MLVA pattern	Accession
0803T57157	Typhimurium	3/11/08	>1600 cases (Outbreak)	Outbreak 1	U292	2	11	13	9	212	JPX.0822.DK	ERR277220
0808S61603	Typhimurium	8/6/08	>1600 cases (Outbreak)	Outbreak 1	U292	2	11	11	9	212	JPX.0411.DK	ERR277226
0902R11254	Typhimurium	2/10/09	>1600 cases (Outbreak)	Outbreak 1	U292	2	11	13	9	212	JPX.0822.DK	ERR277229
000419417	Typhimurium	4/7/00	Background	–	U292	2	11	13	9	212	JPX.0822.DK	ERR274480
0207T641	Typhimurium	7/16/02	Background	–	U292	2	10	16	9	212	JPX.0779.DK	ERR277205
0808F31478	Typhimurium	8/27/08	>200 cases (Outbreak)	Outbreak 2	DT135	2	15	7	10	212	JPX.0855.DK	ERR277223
0903R11327	Typhimurium	3/10/09	>200 cases (Outbreak)	Outbreak 2	DT135	2	15	7	10	212	JPX.0855.DK	ERR277222
0508R6811	Typhimurium	8/24/05	Background	–	DT135	2	11	5	10	212	JPX.0273.DK	ERR277218
0811R10987	Typhimurium	11/28/08	Background	–	DT135	3	18	NA	20	311	JPX.1023.DK	ERR277224
0808R10031	Typhimurium	8/7/08	Background	–	DT135	2	11	11	9	212	JPX.0411.DK	ERR277225
0804R9234	Typhimurium	4/4/08	~ 100 cases (Outbreak)	Outbreak 3	DT3	3	20	7	6	212	JPX.0767.DK	ERR277221
0810R10649	Typhimurium	10/2/08	~ 100 cases (Outbreak)	Outbreak 3	DT3	3	20	7	6	212	JPX.0767.DK	ERR277227
0901M16079	Typhimurium	1/27/09	~ 100 cases (Outbreak)	Outbreak 3	U292	3	20	7	6	212	JPX.0767.DK	ERR277228
0905W16624	Typhimurium	5/15/09	~ 100 cases (Outbreak)	Outbreak 3	DT3	3	14	7	6	212	JPX.1118.DK	ERR277230
0110T17035	Typhimurium	10/30/01	Background	–	DT3	2	11	11	9	212	JPX.0411.DK	ERR277203
0505F37633	Typhimurium	5/13/05	Background	–	DT3	4	15	8	−2	111	JPX.0227.DK	ERR277213
0508R6701	Typhimurium	8/10/05	50 cases. Source: restaurant	Outbreak 4	DT104	3	11	18	17	311	JPX.0253.DK	ERR277214
0508R6707	Typhimurium	8/5/05	50 cases. Source: restaurant	Outbreak 4	NT	3	11	18	17	311	JPX.0253.DK	ERR277216
0508R6762	Typhimurium	8/23/05	50 cases. Source: restaurant	Outbreak 4	DT104	3	11	18	17	311	JPX.0253.DK	ERR277217
0210H31581	Typhimurium	10/24/02	Background	–	DT104	3	14	19	21	311	JPX.1563.DK	ERR277206
0510R6956	Typhimurium	10/19/05	Background	–	DT104	3	12	9	25	311	JPX.1580.DK	ERR277219
0408R5930	Typhimurium	8/26/04	Outbreak	Outbreak 5	DT12	4	4	14	7	211	JPX.0056.DK	ERR277210
0408R5960	Typhimurium	8/24/04	Outbreak	Outbreak 5	DT12	4	4	14	7	211	JPX.0056.DK	ERR277211
0409R5985	Typhimurium	9/8/04	Outbreak	Outbreak 5	DT12	4	4	14	7	211	JPX.0056.DK	ERR277212
0112F33212	Typhimurium	12/21/01	Background	–	DT12	4	13	13	8	211	JPX.0108.DK	ERR277204
0406R5753	Typhimurium	6/30/04	Background	–	DT12	4	17	12	7	211	JPX.0052.DK	ERR277207
0407M287	Typhimurium	7/5/04	Background	–	DT12	4	17	12	7	211	JPX.0052.DK	ERR277208
0407W47858	Typhimurium	7/7/04	Background	–	DT12	4	17	12	7	211	JPX.0052.DK	ERR277209
0508R6706	Typhimurium	8/3/05	Background	–	DT12	4	14	9	10	211	JPX.0167.DK	ERR277215
1004F19825	O:4,12; H:i: –	4/18/10	Outbreak	Outbreak 6	DT120	3	12	10	NA	211	JPX.0005.DK	ERR277232
1005R12913	Typhimurium	5/31/10	Outbreak	Outbreak 6	DT120	3	12	10	NA	211	JPX.0005.DK	ERR277233
1006R12965	Typhimurium	6/16/10	Outbreak	Outbreak 6	DT120	3	12	10	NA	211	JPX.0005.DK	ERR277234
0909R12120	Typhimurium	9/15/09	Background	–	DT120	3	12	9	NA	211	JPX.0007.DK	ERR277231
1007T38029	O:4,5,12; H:i: –	7/12/10	Background	–	DT120	3	14	7	NA	211	JPX.0974.DK	ERR277235
0905R11565	Enteritidis	5/18/09	Outbreak	Enteritidis 1	PT8	–	–	–	–	–	JEG.0001.DK	ERR277236
0905R11609	Enteritidis	5/26/09	Outbreak	Enteritidis 1	PT8	–	–	–	–	–	JEG.0004.DK	ERR277237
0909R12091	Enteritidis	9/4/09	Outbreak	Enteritidis 1	PT8	–	–	–	–	–	JEG.0001.DK	ERR277238
0910R12287	Enteritidis	10/23/09	Background	–	PT8	–	–	–	–	–	JEG.0073.DK	ERR248795

Table 1. Cont.

ID	Serotype	Received date	Outbreak/ Background	Outbreak no.	Phage type	STTR9	STTR5	STTR6	STTR10	STTR3	MLVA pattern	Accession
0909R12018	Enteritidis	9/1/09	Outbreak	Enteritidis 2	PT13a	–	–	–	–	–	JEG.0007.DK	ERR277239
0910R12234	Enteritidis	10/8/09	Outbreak	Enteritidis 2	PT13a	–	–	–	–	–	JEG.0007.DK	ERR277240
0905R11615	Enteritidis	5/29/09	Background	–	PT13a	–	–	–	–	–	JEG.0024.DK	ERR277242
0907R11860	Enteritidis	7/29/09	Background	–	PT13a	–	–	–	–	–	JEG.0021.DK	ERR277243
0807H16988	Derby	7/10/08	Outbreak	Derby outbreak	–	–	–	–	–	–	–	ERR277244
0810W40256	Derby	10/15/08	Outbreak	Derby outbreak	–	–	–	–	–	–	–	ERR277245
0903F3864	Derby	3/11/09	Outbreak	Derby outbreak	–	–	–	–	–	–	–	ERR277246
0807T13477	Derby	7/17/08	Background	–	–	–	–	–	–	–	–	ERR277247
0810F45685	Derby	10/29/08	Background	–	–	–	–	–	–	–	–	ERR277248

doi:10.1371/journal.pone.0087991.t001

BLASTP alignments [31,32]. Any genes having at least 50 percent identity and 50 percent of aligned longest sequence's length (50/50 rule) were considered as a gene cluster [31,33]. The gene clusters that were found in all genomes were collected as a core gene.

SNP Tree

Single nucleotide polymorphisms (SNPs) were identified using a genobox pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [34]. The pipeline consists of various freely available programs. Basically, the paired-end reads from each isolates were aligned against the reference genome, *S.* Typhimurium str. LT2, using Burrows-Wheeler Aligner (BWA) [35]. The average depth coverage was 74. SAMtools [36] 'mpileup' command and bedtools [37] were used to determine and filter SNPs. The qualified SNPs were selected once they met the following criteria: (1) a minimum coverage (number of reads mapped to reference positions) of 20; (2) a minimum distance of 20 bps between each SNP; (3) a minimum quality score for each SNP at 30; and (4) all indels were excluded. The qualified SNPs found within *Salmonella* core genes were ultimately used to make SNP tree because SNPs within the non-core reflect the high proportion of mobile or extra-chromosomal elements, including prophage and genomic islands [14,38].

SNP tree was not only constructed from raw reads but also from contigs or assembled genomes. We used the software package called MUMmer version 3.23 [39]. An application named Nucmer (which is a part of MUMmer) was introduced to align each of contigs to the reference genome. SNPs were determined from the resulting alignments with another MUMmer application called "show-snps" (with options "-CIlrT"). The final set of SNPs was filtered using the following criteria; (1) a minimum distance of 20 bps between each SNP; (2) all indels were excluded.

For each genome, the final qualified SNPs for each genome were concatenated to a single alignment relatively to the position of the reference genome by an in-house perl script. If SNP is not found in the reference genome or the base coverage is less than a minimum setting (20 coverage), it is interpreted as not being a variation and the corresponding base in the reference is expected [34,40]. Subsequently, multiple alignments were employed by MUSCLE from MEGA5 [41]. SNP tree was constructed by MEGA5 using maximum parsimony method [41]. Bootstrapping is frequently used to exhibit the reliability of the branching in a tree. From each sequence, *n* nucleotides are randomly chosen with replacements. These constitute a new set of sequences. A tree is then reconstructed and the tree topology is compared to that of the original one. This procedure of resampling the sites and the subsequent tree reconstruction is repeated 1000 times, and the percentage of times each interior branch is given is noted as bootstrap-value.

Results

The evaluation data consisted of a set of 34 genomes and a set of 47 genomes. The former set contained 34 *S.* Typhimurium strains which 18 isolates were epidemiologically related outbreak strains from 6 different outbreaks, whereas 16 isolates were un-related strains (background or sporadic isolates). The latter set comprised 34 *S.* Typhimurium from the previous set, 8 *S.* Enteritidis of which 5 isolates were outbreak related strains from a couple of outbreaks and 3 were background strains and 5 *S.* Derby of which 3 isolates were outbreak related strains from the same outbreak and 2 isolates were background strains (Table 1).

The performance of typing methods was measured by percentage of concordance. The 100% concordance means all outbreak-related strains from a particular outbreak clustered together and separated from any background isolates.

Traditional *Salmonella* Typing

Pulsed-field gel electrophoresis has been used as a standard procedure for epidemiological outbreak investigations of *Salmonella* [6]. Nonetheless, PFGE gave less discrimination power than WGS typing when applied to closely related strains, e.g strains with the same phage type. Some strains from different outbreaks were grouped together and some outbreak strains were mixed with background isolates (Figure S1).

Whole-genome *Salmonella* Typing

Pan-genome tree. The pan genome tree is the phylogenetic tree based on the profile of presence and absence of genes across genomes [2,26,27]. For the set of 34 genomes, the tree failed to cluster the outbreak strains into the corresponding groups of six different outbreak sources (Figure 1A). The tree only gave the reliable cluster for *S.* Derby outbreak strains (Figure 2A). Additionally, some different outbreak strains were mixed together. This method showed 65% and 64% concordance for the set of 34 and 47 genomes respectively. This is relatively low compared to

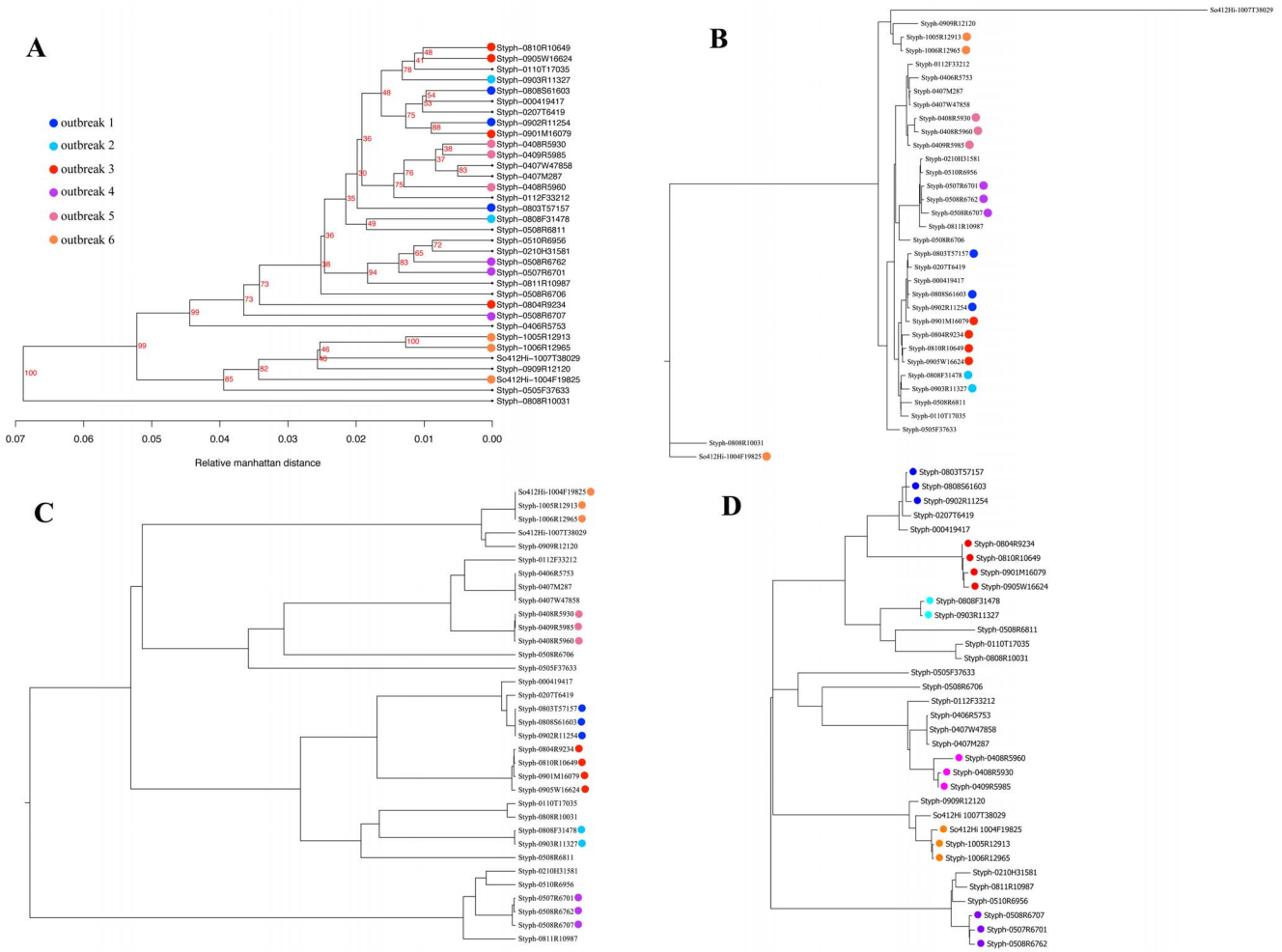


Figure 1. WGS typing results for the set of 34 genomes. (A) pan-genome tree, (B) K-mer tree, (C) nucleotide difference tree and (D) SNP tree. The tested set consists of outbreak-related strains displayed with color label and non-related outbreak strains shown without coloring. The outbreak strains were labeled according to the six different outbreak sources.
doi:10.1371/journal.pone.0087991.g001

the performance from other approaches (Table 2). However, the pan-genome tree revealed high performance for clustering strains according to their phage type (Figure S2).

K-mer tree. K-mer tree was constructed from the frequency profile of k-mers across the selected genomes. The size of k is a sensitive factor for the performance of k-mer tree. A number of various k were evaluated on the set of 34 *S.* Typhimurium. Figure 3 showed an increase in the percentage of concordance with increasing k value. There was a rise in the concordance to a level of 88% concordance at $k = 30$. The percentage remained at this level when $k>30$ suggesting that this range of k achieved the highest performance of k-mer tree. Therefore, we chose $k = 35$ to build the final k-mer tree.

Figure 1B showed that k-mer tree gave higher resolution and more reliable tree than the pan-genome tree. However, some outbreak-related isolates were mixed up with the background strains (Figure 1B). Interestingly, the expanded tree in Figure 2B was capable to place the *S.* Enteritidis outbreak strains into two distinct clusters according to their outbreak groups. The tree also succeeded with clustering *S.* Derby outbreak strains. Nevertheless, the k-mer tree exhibited 88% and 89% concordance for the set of 34 and 47 isolates respectively (Table 2). The time consuming of k-mer tree was only 5.2 minutes per genome (including the time for assemble process). This is the fastest method compared to the others.

Nucleotide difference tree. As a baseline, we implemented a simple approach, the nucleotide difference tree (ND tree), which based on nucleotide difference between a pair of read mapped reference genomes. For the set of 34 *S.* Typhimurium, the ND tree classified outbreak-related strains into six obvious clusters (Figure 1C) with 100% concordance (Table 2). Thus, the typing ability of the ND tree was superior to the pan-genome tree and the k-mer tree. For the set of 47 genomes, the performance of the ND tree was slightly reduced (Figure 2C). The percentage of concordance decreased from 100 to 91% (Table 2).

SNP tree. SNP tree was computed from concatenated qualified SNPs identified from mapping raw reads to core genes of the reference genome [14,38]. From figure 1D, the SNP tree clustered *S.* Typhimurium outbreak-related strains into six clusters with 100% concordance (Table 2) and furthermore differentiated them accurately from the background isolates. For the set of 47 genomes, SNP tree was able to categorized *S.* Derby isolates but unable to ultimately classify the *S.* Enteritidis strains (Figure 2D). The percentage of concordance was dropped from 100 to 91% (Table 2). This is due to the choice of reference genome, SNP tree and ND tree were able to cluster *S.* Enteritidis outbreak strains

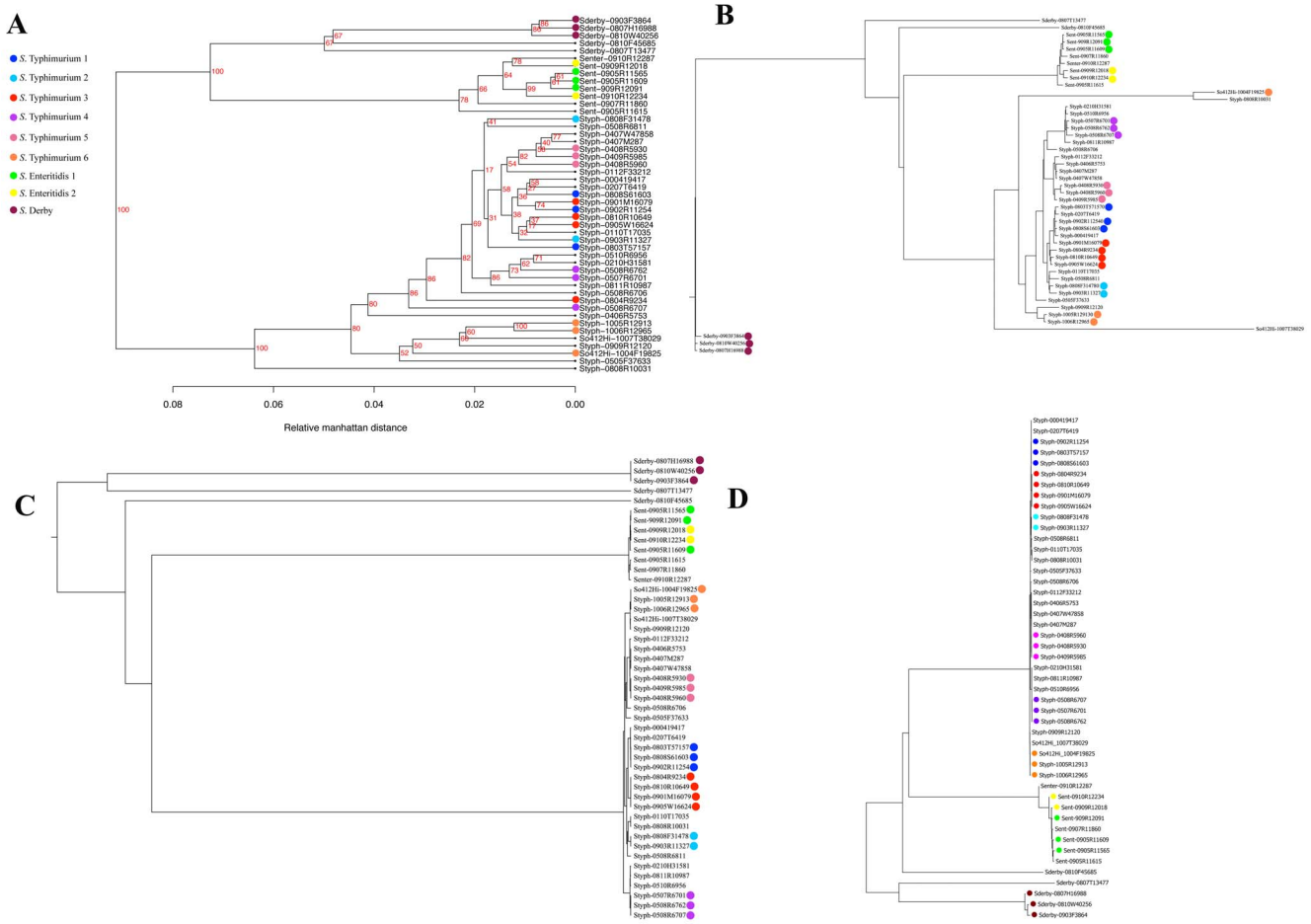


Figure 2. WGS typing results for the set of 47 genomes. (A) pan-genome tree, (B) K-mer tree, (C) nucleotide difference tree and (D) SNP tree. The labeled color was displayed the same as Figure 1.
doi:10.1371/journal.pone.0087991.g002

concordantly by applying publicly available *S.* Enteritis str. P125109 as a reference genome (data not shown). On average, 4.69 Mb of reference genome was covered by *S.* Typhimurium genomes meanwhile the reference genome was mapped with 4.63 Mb and 4.60 Mb when adding *S.* Enteritis and *S.* Derby.

The performance of SNP tree from raw reads was slightly higher than the one from contigs but constructing the SNP tree from contigs was faster (Table 2). In addition, the identified SNPs were distributed thoroughly across core genes of the reference genome (Figure 4) suggesting that the mutation occurred randomly through the core genes.

Figure 5 revealed that minimum and maximum number of SNP difference within the outbreak strains were significantly less than those numbers between outbreak-related isolates and background isolates. The number of SNP difference between isolates within outbreaks ranged from 2 to 12 except the outbreak 5 (DT12) where the maximum number was relatively high (3–30 SNPs). Besides, the number of days within outbreak strains was unrelated

Table 2. Evaluation results.

WGS typing methods	Percentage of concordance		Time (Minutes per genome)	Reference based method	Type of input
	34 isolates	47 isolates			
Pan-genome tree	65	64	13	Reference free	Contigs
K-mer tree	88	89	5.2	Reference free	Contigs
Nucleotide difference tree	100	91	15	Reference-based	Raw reads
SNP tree (raw reads)	100	91	20	Reference-based	Raw reads
SNP tree (contigs)	100	89	5.5	Reference-based	Contigs

doi:10.1371/journal.pone.0087991.t002

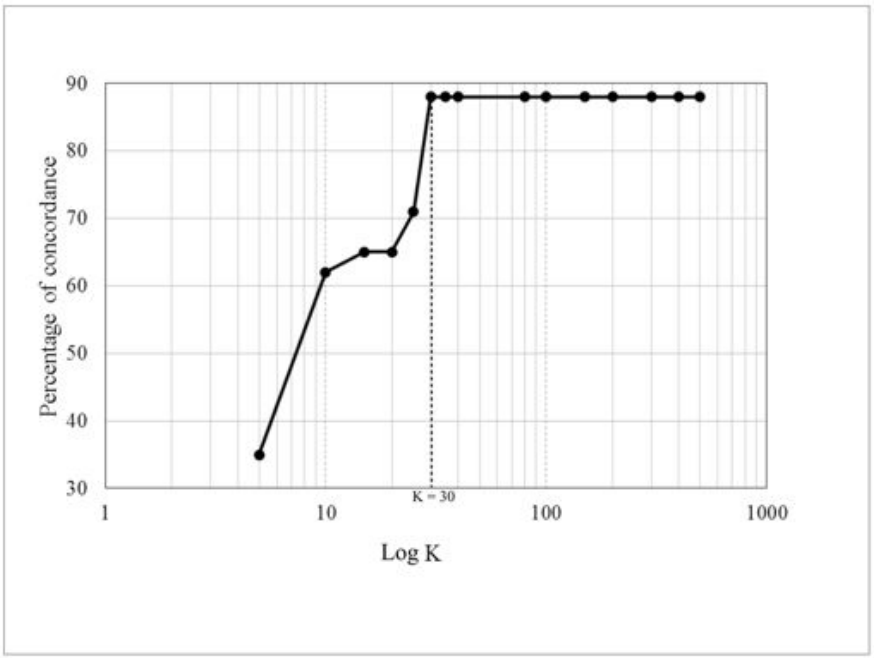


Figure 3. Percentage of concordance of k-mer tree on various size of k. This evaluation was conducted on the set of 34 *S*. Typhimurium.
doi:10.1371/journal.pone.0087991.g003

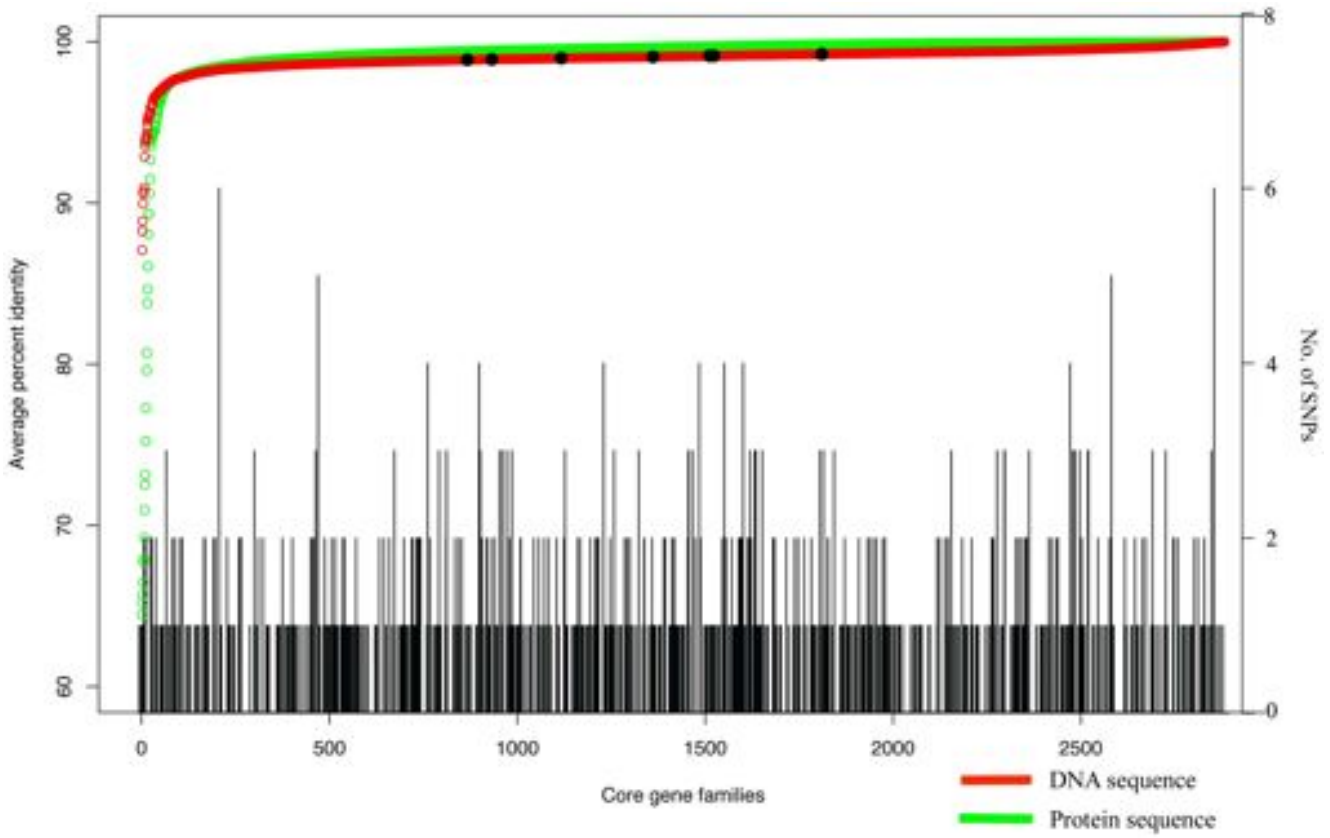


Figure 4. Distribution of SNPs across *Salmonella* core genes. Black bars represent number of SNPs at each core gene. Red and green small circles are core genes in the form of DNA and protein sequences respectively. The seven black dots represent house-keeping genes for MLST analysis of *Salmonella*.
doi:10.1371/journal.pone.0087991.g004

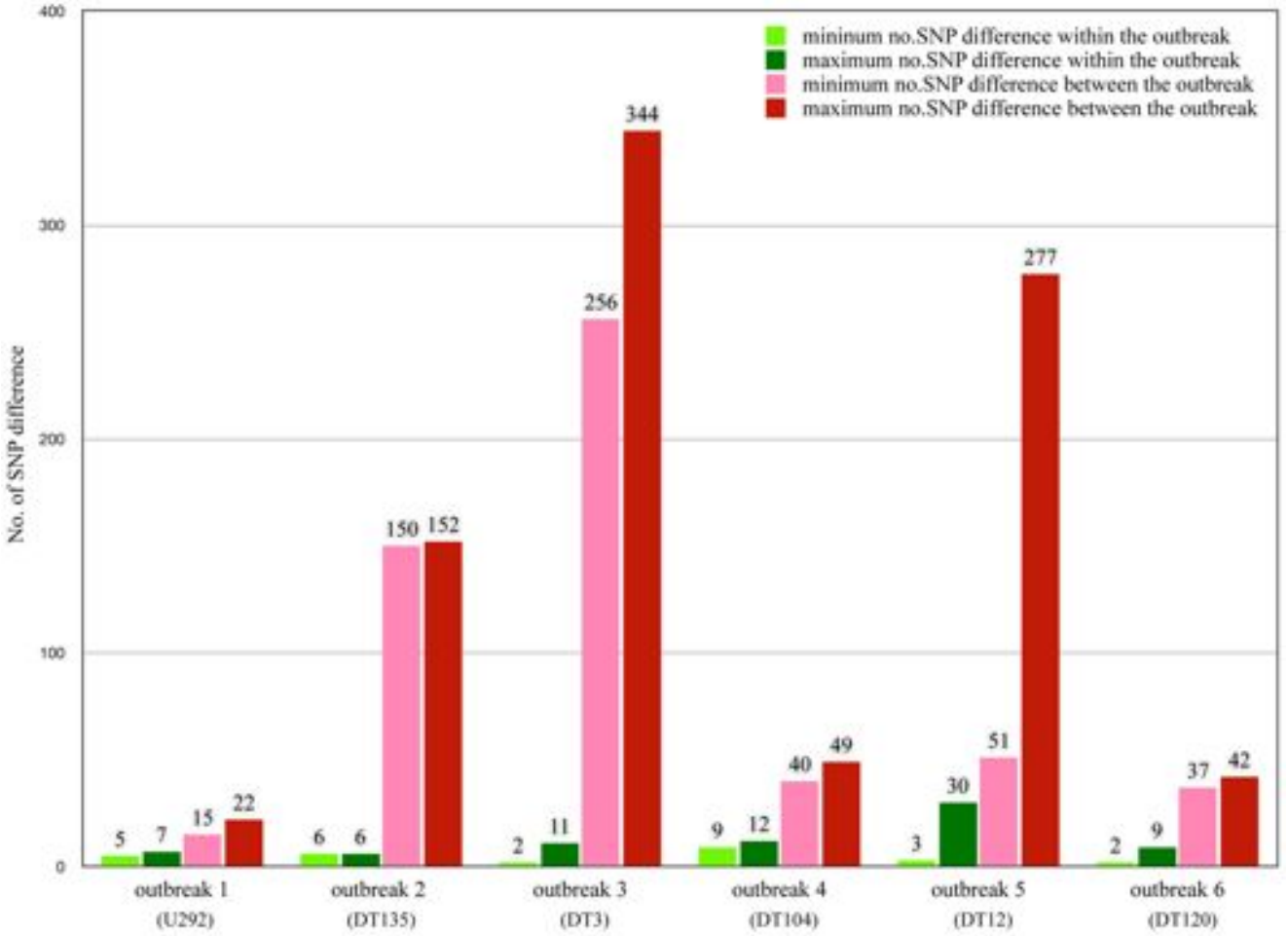


Figure 5. Minimum and maximum number of SNP difference. Green shaded bars show the minimum and maximum number of SNP difference between isolates within outbreaks and red shaded bars represent the number of SNP difference between outbreak-related isolates and background isolates.
doi:10.1371/journal.pone.0087991.g005

to the number of SNP difference (Figure S3) and this relation seems to be random.

Comparison with Published Studies

Four publicly available *Salmonella* outbreak dataset were integrated and analyzed by SNP approach. These data comprised of background and outbreak-related strains except *S.* Heidelberg that contained only outbreak strains. An average number of SNP difference or pairwise SNP distance between strains within outbreaks and between outbreak-related strains and background strains were summarized in Figure 6. *S.* Montevideo and *S.* Enteritidis supported our finding that a SNP distance within outbreak strains was less than that between outbreak and background strains. Interestingly, *S.* Agona showed the higher number of SNP difference within outbreak strains and these numbers from two sub-outbreak clusters were higher than the SNP distance between background and outbreak strains. The number of SNP differences between strains within an outbreak is likely to vary for each serotype making it difficult to find the threshold for the case definition of an outbreak.

We reproduced SNP tree and k-mer tree based on 271 genomes from publicly available *Salmonella* genomes together with the genomes under study (Figure S4A and S4B). It was not possible to reproduce the tree by ND tree because most of the published data are assembled genomes and the ND tree was invented primarily for raw reads. The reproduced trees from SNP and k-mer formed distinct clusters according to serotypes. However, combining different serovar strains, k-mer and SNP trees illustrated the similar tree topology of *S.* Typhimurium cluster as they showed in Figure 1B and 1D respectively. Nonetheless, the reproduced SNP tree exhibited less resolution than the tree constructed from the strains with identical serovar as in Figure 1D.

Discussions

The objective of this study was to determine the strengths and drawbacks of WGS using different analytic approaches compared to traditional typing method, PFGE, for retrospectively outbreak typing of *Salmonella*. A set of thirty-four human *S.* Typhimurium strains from six different outbreaks together with background strains plus eight *S.* Enteritidis isolates from two outbreaks and five *S.* Derby strains from a single outbreak were used as test sets. A number of recent studies have already used WGS for epidemiological typing of single outbreaks [13,14,17]. However, these studies have only used SNP analysis and not other analytic procedures. We evaluated different of analytical approaches on the WGS data set and compared to PFGE typing - the gold standard

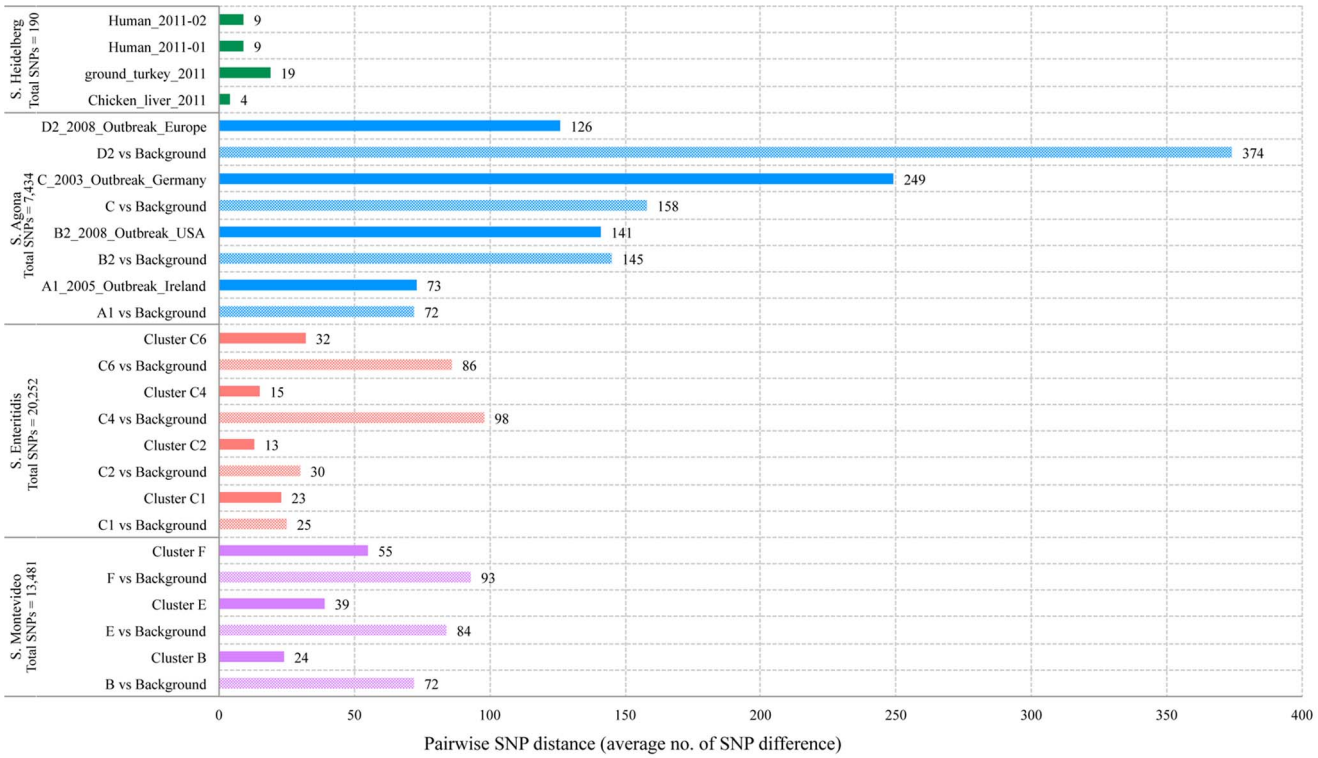


Figure 6. The pairwise SNPs distance. This is the average number of SNP difference between strains within outbreaks and between outbreak-related strains and background strains from the four published dataset.
doi:10.1371/journal.pone.0087991.g006

method for epidemiological studies. In our study, WGS based typing using SNP tree and ND tree was able to compete with PFGE for outbreak clustering.

The performance of the four selected WGS based typing methods was validated based on the outbreak related *Salmonella enterica* strains. Pan-genome tree failed to perform accurate clusters as the variation in protein level among the outbreak strains was not appropriate for outbreak typing, although the pan-genome tree showed meaningful clusters corresponding to phage types. This could be due to the content of prophages. The k-mer tree gave the expected clustering but was still unable to employ the complete outbreak typing. Interestingly, the k-mer tree revealed a better clustering when combining *Salmonella* strains from different serovars. This is most likely because the k-mer tree is independent from the reference genome. Another advantage of k-mer analysis is that the frequencies-based approach is much faster. Thus, it is expected to be applicable for both closely and more distantly related strains with very short time consumption for analysis. On the other hand, a deficiency is the loss of information as the huge amount of DNA sequence data is condensed into a vector of k-mer counts. Furthermore, The order of k-mers in compared sequences is neglected [30]. The nucleotide difference tree (ND tree) identified the number of nucleotide difference between a pair of raw read mapped reference genomes rather than identify the difference as SNP. This method gave the results similarly to the SNP tree. Additionally, it is important to note that SNP not being found in the reference genome is considered as not being a variation and the corresponding nucleotide from the reference is expected. This might not always be the right choice. The ND tree does not face this problem, as it does not require the concatenated sequence for alignment. ND tree was found to be somewhat sensitive to its setting. In initial calculations the mismatch score was set to -1, and in this tree all *S.* Enteritidis and *S.* Derby strains became identical (data not showed). The final results used a mismatch score as -3, which is also the default in the short read alignment program, BWA.

Ultimately, SNP and ND trees were equally superior methods for clustering outbreak related isolates of *S.* Typhimurium (Figure 1C and 1D). As mentioned above, ND tree was sensitive to the parameter settings, while SNP tree failed to categorize strains with different serovars because this method depends heavily on the reference genome and this has to be closely related to the strains investigated for example the reference genome should be at least the same serovar as the strains under study. Using an inappropriate reference genome will cause exceed number of SNPs which affects the final SNP tree for instance the decreasing of the percentage concordance when adding strains with different serovars from the reference genome (Table 2, SNP tree with a set of 47 genomes). In addition, SNP tree constructed from contigs exhibited slightly less concordance than the one from the raw reads. In term of speed, the SNP tree from contigs can be achieved very fast (almost as fast as k-mer tree). It might be an alternative choice of using SNP tree for real-time typing.

We found that the numbers of SNP difference between isolates within outbreaks were very small and ranged from 2 to 12 with an exception for the outbreak 5 (DT12) where the number ranged from 3 to 30 SNP differences. Comparing to publicly available *Salmonella* genomes, the SNP distance between strains within outbreaks was possibly ranged from 4 to 249 depending on serotype suggesting that finding a general threshold to define an outbreak for all *Salmonella* might not be possible. However, these numbers may be useful as an indicator of expected SNP distance in a particular serovar or a sub-outbreak cluster within serovar. Nevertheless, by using a small number of isolates from specific

outbreaks, this reduced sampling may be introduce some of other variables affecting the predictions. It may take dozens of isolates to determine the actual scope or threshold of an outbreak.

Recent studies support SNP tree as an outbreak surveillance tool such as *S.* Montevideo outbreak in United States [17,42], *S.* Enteritidis shell egg outbreak in US in 2010 [25], *S.* Agona [24] and a 2011 multistate outbreak in the US of *S.* Heidelberg [22,23]. Nonetheless, the SNP detection and validation need to be improved, and this method needs to be further evaluated in other bacterial pathogens to elucidate the usefulness of using SNP tree. Perhaps, for further pathogens, other approaches might be the most superior beside SNP analysis. In addition, it is especially a need to determine the importance of using different sequencing platforms, different analytic procedures and different reference strains for creating the SNP trees. Moreover, the robustness of this analytical approach for cluster detection in a routine setting has to be evaluated. The fact that the tree topology may give less resolution when new strains are added might cause some problems in the interpretation in a routine setting and over time.

In our study, we were unable to find an association between time (days) of isolation and number of SNP difference between isolates belonging to the same outbreak. This contrasts studies of methicillin-resistant *Staphylococcus aureus* (MRSA) spreading between humans in hospital community, where the time and number of SNPs are correlated [14]. This might be due to the dissimilarities in the epidemiology of these bacterial pathogens. MRSA transfers from human to human within a hospital, whereas *Salmonella* has its natural reservoir in various sources, animals and human. Thus, the transmission route of *Salmonella* to human is indirect and even though two strains are isolated with a given time interval this might not entirely reflect the number of generations that they differ. Nonetheless, this observation is in agreement with that was reported by Okoro *et al* [43]. They show that the number of days (23–486 days) between isolation of index and recurrent isolates of *S.* Typhimurium from infected patients had no obvious impact on the numbers of SNP differences accumulated, and suggest the existence of groups of isolates that comprise single clonal haplotypes with virtually no genetic change over time.

The strains included in this study were selected based on detailed epidemiological information as estimated to belong or not belonging to the same outbreak. Since the true epidemiology is not known, it cannot be excluded that strains not being part of an outbreak have been falsely included or that true outbreak strains have been falsely categorized as non-outbreak related. Based on the detailed epidemiological information available and carefully selection of isolates, we do believe that the reference material reflects the true epidemiology and that the methods SNP and ND are superior to the currently used methods for epidemiological typing such as PFGE. However, only time and routine implementation of the new WGS technologies in routine investigations will provide the value of WGS as supporting outbreak detection and control.

It is also important to note that WGS is as all other typing tools to support for decision making and should always be used in combination with epidemiological and/or clinical information. For example, the different phylogenetic trees shown in this study were not meaningful without any support from epidemiological information (the color dots in the trees). Thus, it is essential to combine epidemiological data and whole genome sequencing results.

In conclusion, this study suggests that WGS and analysis using SNP and/or nucleotide difference approaches are superior methodologies for epidemiological typing of *S.* Typhimurium isolates and might be very successfully applied for outbreak detection. For the very fast but rough result, k-mer tree might meet this requirement with constructing the tree in high speed and giving high accuracy in clade level.

Supporting Information

Figure S1 An UPGMA band based comparison of pulsed-field gel electrophoresis (PFGE) *XbaI* profiles.
(PDF)

Figure S2 Pan-genome tree with phage typing labels.
(PDF)

Figure S3 The relation between number of days and number of SNP difference among the outbreak strains.
(PDF)

Figure S4A SNP tree constructed from 271 genomes from published data and *Salmonella* genomes under this study.
(PDF)

Figure S4B K-mer tree constructed from 271 genomes from published data and *Salmonella* genomes under this study.
(PDF)

Author Contributions

Conceived and designed the experiments: PL FMA. Performed the experiments: PL EMN RSK OL. Analyzed the data: PL. Contributed reagents/materials/analysis tools: PL EMN RSK OL. Wrote the paper: PL EMN OL FMA.

References

1. Hohmann EL (2001) Nontyphoidal salmonellosis. Clin Infect Dis 32: 263–269.
2. Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup FM, Ussery DW (2012) Genomic variation in Salmonella enterica core genes for epidemiological typing. BMC genomics 13: 88.
3. Fisher I (1999) Salmonella enteritidis in Western Europe 1995–98 - a surveillance report from Enter-net. Euro Surveill 4: 56.
4. Didelot X, Bowden R, Wilson DJ, Peto TE a, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nature reviews Genetics 13: 601–612.
5. Pallen MJ, Loman NJ, Penn CW (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Current opinion in microbiology 13: 625–631.
6. Foley SL, Zhao S, Walker RD (2007) Comparison of Molecular Typing Methods for the Differentiation of Salmonella Foodborne Pathogens. Foodborne Pathog Dis 4: 253–276.
7. Dewaele I, Rasschaert G, Bertrand S, Wildemauwe C, Wattiau P, et al. (2012) Molecular characterization of Salmonella Enteritidis: comparison of an optimized multi-locus variable-number of tandem repeat analysis (MLVA) and pulsed-field gel electrophoresis. Foodborne pathogens and disease 9: 885–895.
8. Campioni F, Davis M, Medeiros MIC, Falcão JP, Shah DH (2013) MLVA typing reveals higher genetic homogeneity among S. Enteritidis strains isolated from food, humans and chickens in Brazil in comparison to the North American Strains. International journal of food microbiology 162: 174–181.
9. Petersen RF, Litrup E, Larsson JT, Torpdahl M, Sørensen G, et al. (2011) Molecular Characterization of Salmonella Typhimurium Highly Successful Outbreak Strains. Foodborne Pathog Dis 8: 655–661.
10. Torpdahl M, Sørensen G, Lindstedt B-A, Nielsen EM (2007) Tandem repeat analysis for surveillance of human Salmonella Typhimurium infections. Emerging infectious diseases 13: 388–395.
11. Wilson DJ (2012) Insights from genomics into bacterial pathogen populations. PLoS pathogens. doi: 10.1371/journal.ppat.1002874.
12. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, et al. (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS pathogens. doi: 10.1371/journal.ppat.1002824.
13. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, et al. (2011) Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio. doi:10.1128/mBio.00157-11.

14. Harris SR, Feil EJ, Holden MTG, Quail M a, Nickerson EK, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327: 469–474.
15. Okoro CK, Kingsley R a, Connor TR, Harris SR, Parry CM, et al. (2012) Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. Nature genetics 44: 1215–1221.
16. Dunne WM, Westblade LF, Ford B (2012) Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. Eur J Clin Microbiol Infect Dis 31: 1719–1726.
17. Allard MW, Luo Y, Strain E, Li C, Keys CE, et al. (2012) High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC genomics 13: 32.
18. Hendriksen RS, Le Hello S, Bortolaia V, Pulsrikarn C, Nielsen EM, et al. (2012) Characterization of isolates of Salmonella enterica serovar Stanley, a serovar endemic to Asia and associated with travel. J Clin Microbiol 50: 709–720.
19. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 18: 821–829.
20. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, et al. (2012) Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 50: 1355–1361.
21. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, et al. (2012) Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67: 2640–2644.
22. Hoffmann M, Luo Y, Lafon PC, Timme R, Allard MW, et al. (2013) Genome Sequences of Salmonella enterica Serovar Heidelberg Isolates Isolated in the United States from a Multistate Outbreak of Human. 1: 1–2. doi:10.1128/genomeA.00004-12.
23. Hoffmann M, Zhao S, Luo Y, Li C, Folster JP, et al. (2012) Genome sequences of five Salmonella enterica serovar Heidelberg isolates associated with a 2011 multistate outbreak in the United States. Journal of bacteriology 194: 3274–3275.
24. Zhou Z, McCann A, Litrup E, Murphy R, Cormican M, et al. (2013) Neutral genomic microevolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS genetics. doi: 10.1371/journal.pgen.1003471.
25. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, et al. (2013) On the evolutionary history, population genetics and diversity among isolates of Salmonella Enteritidis PFGE pattern JEGX01.0004. PloS one. doi: 10.1371/journal.pone.0055254.
26. Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. Standards in genomic sciences 2: 135–141.
27. Vesth T, Lagesen K, Acar Ö, Ussery D (2013) CMG-biotools, a free workbench for basic comparative microbial genomics. PloS one. doi: 10.1371/journal.pone.0060120.
28. Cheng J, Cao F, Liu Z (2013) AGP: A Multimethods Web Server for Alignment-Free Genome Phylogeny. Molecular biology and evolution. doi: 10.1093/molbev/mst021.
29. DeSantis TZ, Keller K, Karaoz U, Alekseyenko AV, Singh NNS, et al. (2011) Simrank: Rapid and sensitive general-purpose k-mer search tool. BMC ecology. doi: 10.1186/1472-6785-11-11.
30. Yu H-J (2013) Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. Gene 518(2): 419–24.
31. Ussery D, Wassenaar T, Borini S (2008) Computing for Comparative Genomics: Bioinformatics for Microbiologists (Computational Series). London: Springer Verlag.
32. Friis C, Wassenaar TM, Javed M a, Snipen L, Lagesen K, et al. (2010) Genomic characterization of Campylobacter jejuni strain M1. PloS one. doi: 10.1371/journal.pone.0012253.
33. Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C (2011) The Salmonella enterica pan-genome. Microbial ecology 62: 487–504.
34. Leekitcharoenphon P, Kaas RS, Thomsen MCF, Friis C, Rasmussen S, et al. (2012) snpTree - a web-server to identify and construct SNP trees from whole genome sequence data. BMC genomics 13 Suppl 7: S6.
35. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25: 1754–1760.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
37. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) 26: 841–842.
38. Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, et al. (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS pathogens. doi: 10.1371/journal.ppat.1002129.
39. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research 30: 2478–2483.
40. Leekitcharoenphon P, Friis C, Zankari E, Svendsen CA, Price LB, et al. (2013) Genomics of an emerging clone of Salmonella serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo. J Infect Dev Ctries 7: 696–706.
41. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
42. Gieraltowski L, Julian E, Pringle J, Macdonald K, Quilliam D, et al. (2012) Nationwide outbreak of Salmonella Montevideo infections associated with contaminated imported black and red pepper: warehouse membership cards provide critical clues to identify the source. Epidemiology and infection 141(6): 1244–52.
43. Okoro CK, Kingsley R a, Quail M a, Kankwatira AM, Feasey N a, et al. (2012) High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal salmonella typhimurium disease. Clinical infectious diseases 54: 955–963.

Article III

snpTree--a web-server to identify and construct SNP trees from whole genome sequence data.

ARTICLE III

Pimlapas Leekitcharoenphon,[1,2*] Rolf S. Kaas,[1,2] Martin Christen Frølund Thomsen,[2] Carsten Friis,[1] Simon Rasmussen,[2] Frank M. Aarestrup,[1]

[1] Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs Lyngby, Denmark

*Corresponding author: **Pimlapas Leekitcharoenphon,**
Division for Epidemiology and Microbial Genomics,
National Food Institute, Technical University of Denmark,
Kgs. Lyngby, Denmark
E-mail: pile@food.dtu.dk

BMC Genomics 2012;13 Suppl 7:S6.

Leekitcharoenphon et al. BMC Genomics 2012, 13(Suppl 7):S6
http://www.biomedcentral.com/1471-2164/13/S7/S6

BMC Genomics

PROCEEDINGS Open Access

snpTree - a web-server to identify and construct SNP trees from whole genome sequence data

Pimlapas Leekitcharoenphon[1,2*], Rolf S Kaas[1,2], Martin Christen Frølund Thomsen[2], Carsten Friis[1], Simon Rasmussen[2], Frank M Aarestrup[1]

From Asia Pacific Bioinformatics Network (APBioNet) Eleventh International Conference on Bioinformatics (InCoB2012)
Bangkok, Thailand. 3-5 October 2012

Abstract

Background: The advances and decreasing economical cost of whole genome sequencing (WGS), will soon make this technology available for routine infectious disease epidemiology. In epidemiological studies, outbreak isolates have very little diversity and require extensive genomic analysis to differentiate and classify isolates. One of the successfully and broadly used methods is analysis of single nucletide polymorphisms (SNPs). Currently, there are different tools and methods to identify SNPs including various options and cut-off values. Furthermore, all current methods require bioinformatic skills. Thus, we lack a standard and simple automatic tool to determine SNPs and construct phylogenetic tree from WGS data.

Results: Here we introduce snpTree, a server for online-automatic SNPs analysis. This tool is composed of different SNPs analysis suites, perl and python scripts. snpTree can identify SNPs and construct phylogenetic trees from WGS as well as from assembled genomes or contigs. WGS data in fastq format are aligned to reference genomes by BWA while contigs in fasta format are processed by Nucmer. SNPs are concatenated based on position on reference genome and a tree is constructed from concatenated SNPs using FastTree and a perl script. The online server was implemented by HTML, Java and python script.
The server was evaluated using four published bacterial WGS data sets (*V. cholerae*, *S. aureus* CC398, *S.* Typhimurium and *M. tuberculosis*). The evalution results for the first three cases was consistent and concordant for both raw reads and assembled genomes. In the latter case the original publication involved extensive filtering of SNPs, which could not be repeated using snpTree.

Conclusions: The snpTree server is an easy to use option for rapid standardised and automatic SNP analysis in epidemiological studies also for users with limited bioinformatic experience. The web server is freely accessible at http://www.cbs.dtu.dk/services/snpTree-1.0/.

Background

The dramatic decrease in cost for whole-genome sequencing (WGS) has made this technology economically feasible as a routine tool for scientific research, including infectious disease epidemiology. In addition, WGS has major applications for health service providers working with infectious diseases [1] as such to deliver high-resolution genomic epidemiology as the ultimate typing method for bacteria.

The ideal microbial typing technique should enable differentiation of epidemiological unrelated strains and group epidemiological related (outbreak) strains, [2] and give information that will help to understand the evolutionary history of multiple strains within a clonal lineage [1,2]. Although some current technologies are highly informative like MLST or PFGE, they have limited resolution when applied to closely related isolates and different methods often have to be applied in different situations [1,2].

* Correspondence: pile@food.dtu.dk
[1]National Food Institute, Building 204, Technical University of Denmark, 2800 Kgs Lyngby, Denmark 4444
Full list of author information is available at the end of the article

Especially outbreak isolates normally have very little diversity and require extensive genomic methods to differentiate and catagorize the isolates [3]. Single nucleotide polymorphisms (SNPs) also show relatively low mutation rates and are evolutionarily stable. Moreover, SNPs analysis has successfully been used for determining broad patterns of evolution in many recent studies [4-6].

Currently, There are a number of available non-commercial NGS genotype analysis software such as SOAP2 [7], GATK [8] and SAMtools [9]. Nonetheless, all of the software require bioinformatic skills, various options, various setting and they do not have a user friendly web-interface.

Here we introduce snpTree. A server for online-automatic SNP analysis and SNP tree construction from sequencing reads as well as from assembled genomes or contigs. The server is a pipeline which intregrates avaliable SNPs analysis softwares such as SAMtools [9] and MUMmer [10], with customized scripts. The performance of the server was evaluated with four published bacterial WGS data set; *Vibrio cholerae* [3], *Staphylococcus aureus* CC398 [6], *Salmonella* Typhimurium [11] and *Mycobacterium tuberculosis* [12].

Implementation

The snpTree server was created to handle both WGS data and assembled genomes to generate a phylogenetic tree based on SNPs data. The overall process is shown in Figure 1. For raw reads (Figure 1A), snpTree use an in-house toolbox (Genobox) for mapping and genotyping which consists of avaliable programs for next-generation sequencing analysis such as Burrows-Wheeler Aligner, BWA [13] and software package for SNPs calling and genotyping, SAMtools [9]. The source code of Genebox is available at https://github.com/srcbs/GenoBox. For contigs or assembled genomes (Figure 1B), MUMmer [10] is used for both reference genome alignment and SNPs identification processes.

The web-server contains more than 2,000 completed reference genomes collected from NCBI Genome database (accessed on April 2012).

SNPs identification from WGS

Prior to mapping raw reads to a proper reference genome, the sequence data in fastq format are filtered and trimmed according to the following criteria [14]: (i) reads with N's are removed, (ii) if a read matches a minimum of 25 nt of a sequencing primer/adaptor the reads are trimmed at the 5' coordinate of match, (iii) the 3' tail bases are trimmed if the quality score is less than 20, (iv) the minimum average quality of the read should be 20 and the read length after trimming should be at least 20 nt.

Trimmed raw reads are aligned against a reference genome using BWA [13] with minimum mapping quality equal to 30 as a default (Figure 1A). BWA is based on an effective data compression algorithm called Burrows-Wheeler transform (BWT) that is fast, memory-efficient and espectially useful for aligning short reads [15].

SNPs calling and filtering are accomplished by SAMtools that is a software package for parsing and manipulating alignments in the generic alignment format (SAM/BAM format) [9]. The snpTree server allows users to set a couple of parameters to filter SNPs, a minimum coverage and a minimum distance between each SNPs (prune). The default for both cut-offs is set to 10 and additionally all heterozygous SNPs are filtered because these are likely mapping errors in haploid chromosomes. The identifed SNPs are concluded into a VCF file.

SNPs identification from assembled genomes

A pipeline has been developed around the software package MUMmer version 3.23 [10] (Figure 1B). An application named Nucmer, which is part of MUMmer, is used to align each of *de novo* assemblies to a reference genome chosen by the user (default settings). SNPs are then called from the resulting alignments with another MUMmer application named "show-snps" (with options "-CIlrT"). A pruning is then applied, if chosen by the user, and the SNPs are written into a VCF formatted file for each of the analyzed genomes.

SNPs tree construction

One VCF formatted file is needed for each Operational Taxonomic Unit (OTU). The SNPs are then concatenated into a single alignment by ignoring indels. Including indels would disturb the position of SNPs in the sigle alignment. To include indels in any trees, it requires some sensible way to represent them numerically as distances in an evolutionary space, and there is no any ways to achieve this. Indels could theoretically be included in a multiple sequence alignment, since such alignments can handle gaps but it's difficult to score them. "Blast-like" gap penalties certainly would not work, since they are optimized for much larger gaps, e.g. recombination events.

It is important to note that SNPs not found in a VCF file is interpreted as not being a variation and the corresponding base in the reference is expected. This might not always be the right choice, because a SNP not found in a VCF file could be a result of an INDEL. It is expected to be a rare case and probably won't disturb the phylogenetic signal.

The alignment is passed on to Fastree [16], which creates a maximum likelihood tree from the SNP alignment.

snpTree server output

snpTree server provides an output to users with SNPs tree figure in SVG format, number of SNPs and other relevant output files such as (i) SNPs files, which contains

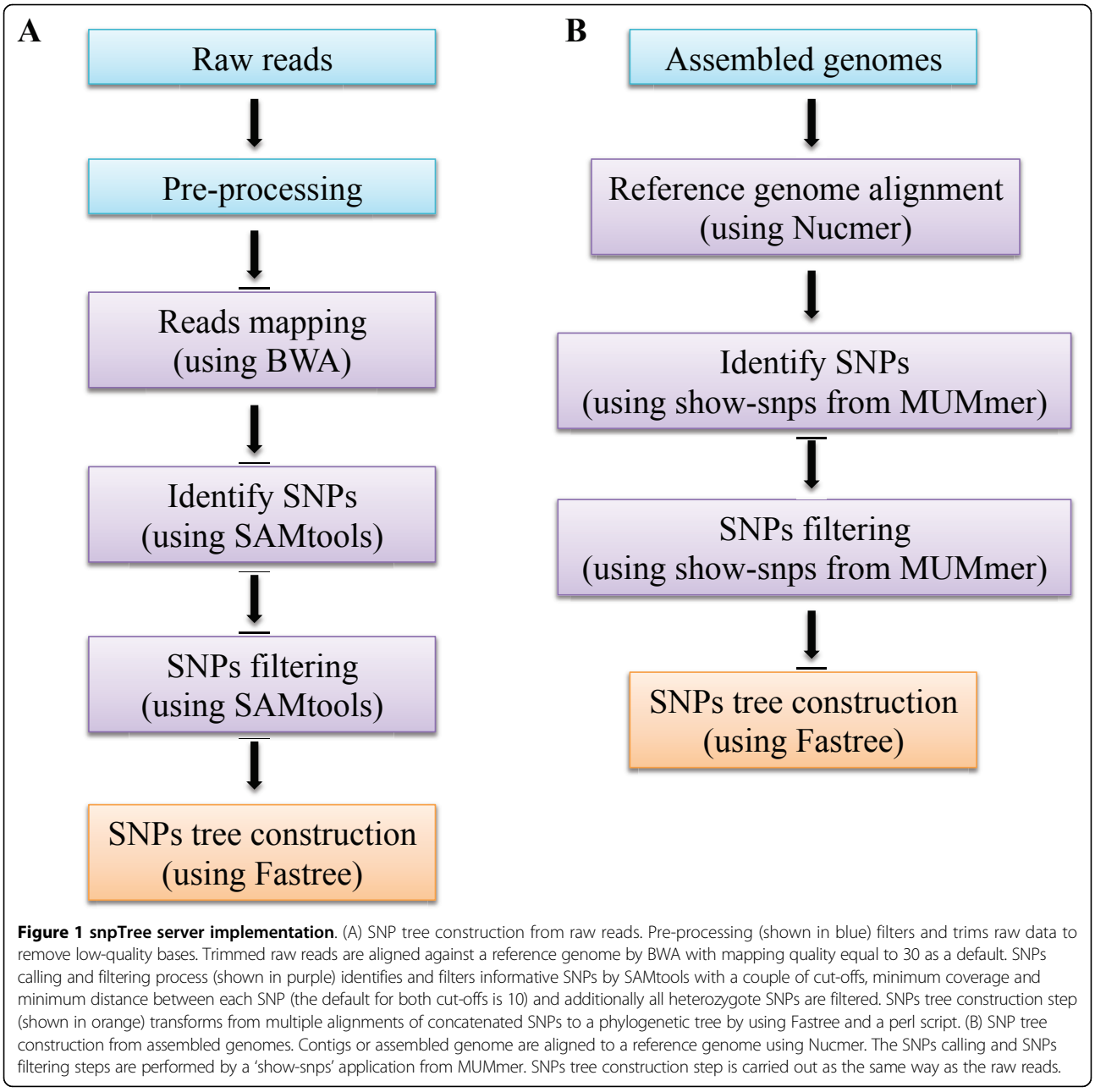


Figure 1 snpTree server implementation. (A) SNP tree construction from raw reads. Pre-processing (shown in blue) filters and trims raw data to remove low-quality bases. Trimmed raw reads are aligned against a reference genome by BWA with mapping quality equal to 30 as a default. SNPs calling and filtering process (shown in purple) identifies and filters informative SNPs by SAMtools with a couple of cut-offs, minimum coverage and minimum distance between each SNP (the default for both cut-offs is 10) and additionally all heterozygote SNPs are filtered. SNPs tree construction step (shown in orange) transforms from multiple alignments of concatenated SNPs to a phylogenetic tree by using Fastree and a perl script. (B) SNP tree construction from assembled genomes. Contigs or assembled genome are aligned to a reference genome using Nucmer. The SNPs calling and SNPs filtering steps are performed by a 'show-snps' application from MUMmer. SNPs tree construction step is carried out as the same way as the raw reads.

identified SNPs including indels for each input genome in VCF format [17], (ii) cancatenated SNPs in newick, phylip and fasta format, (iii) SNPs annotation files which give users an overview of nucleotide changes or amino acid changes from SNPs including which input genomes contain which SNPs as well as information about synonymous and non-synonymous SNPs (Additional file 1). An example of output is shown in Figure 2.

Results and discussion

The snpTree was evaluated using raw reads and assembled genomes from four published bacterial WGS data sets (*V. cholerae* [3], *S. aureus* CC398 [6], *S.* Typhimurium [11] and *M. tuberculosis* [12]). The evaluation was considered based on tree topology as well as the reference genome's position of identifed SNPs.

Evaluation of tree topology and SNPs position

WGS from published data set were subjected to snpTree server in order to generate SNP trees. The tree topology evaluation was based on percentage of concordance. If the strain in the tree from snpTree server matches exactly with the tree from published data, it was considered as an exact match. If the strains were grouped into

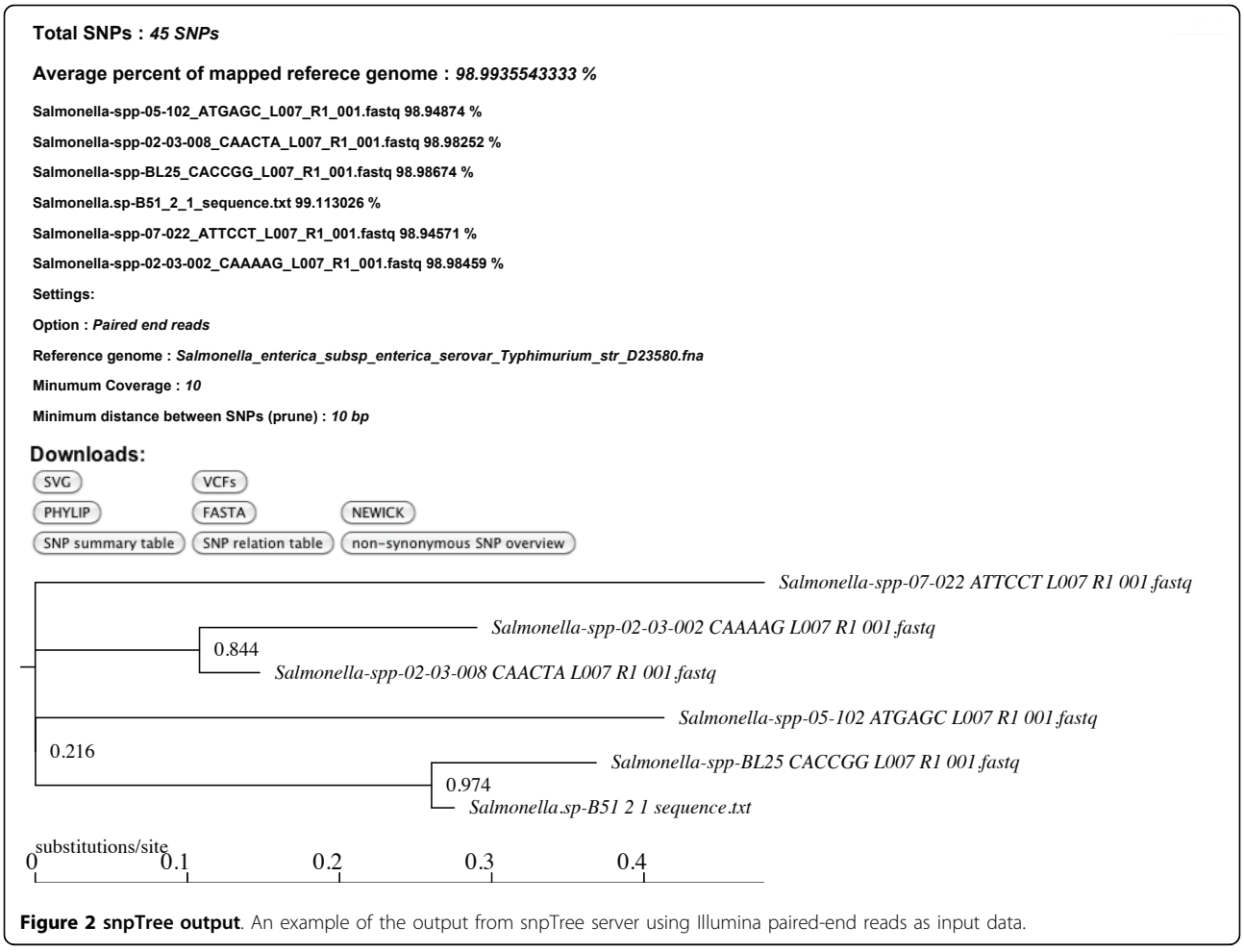


Figure 2 snpTree output. An example of the output from snpTree server using Illumina paired-end reads as input data.

the same cluster with published data, it was considered as a cluster match. In addition, the snpTree server was evaluated with assembled genomes or contigs. The raw reads were assembled prior by *de novo* assembly using Velvet 1.1.04 [18]. The assembled genomes were processed to snpTree server to make SNP trees.

V. cholerae data set

The evaluation results are summarized in Table 1. For the *V. cholerae* data set, the performance of snpTree from raw reads (Figure 3) and contigs (Additional file 2) were accurate in term of exact match and cluster match. From Figure 3, all of genomes were grouped into the same clusters as in the original tree. In the Nepal-1 cluster, there are only 3 genomes that are not in the same position compared to the original tree. However, the isolates in Nepal-1 group are highly homogeneous and there are some synapomorphic SNPs (genome position that has mutated the new nucleotide which shared with all descendants) supporting its unique identities [3].

The percentage of overlapped and non-overlapped SNPs between published data and snpTree server is illustrated in Figure 4A for raw reads and Figure 4B for assembled genomes. For *V. cholera*, both raw reads and contigs (Figure 4), the snpTree server identified SNPs mostly from the same position in published data (95%

Table 1 Evaluation table

Data set	Percentage of concordance	
	Exact match	cluster match
V. cholerae (raw reads)	91	100
***V. cholerae* (contigs)**	**85**	**100**
S. aureus CC398 (raw reads)	88	96
***S. aureus* CC398 (contigs)**	**87**	**97**
S. typhimurium (raw reads)	61	100
***S. typhimurium* (contigs)**	**53**	**100**
M. tuberculosis (raw reads)	58	78
***M. tuberculosis* (contigs)**	**25**	**72**

The percentage of concordance from comparing SNP trees from snpTree server against the four published data set.

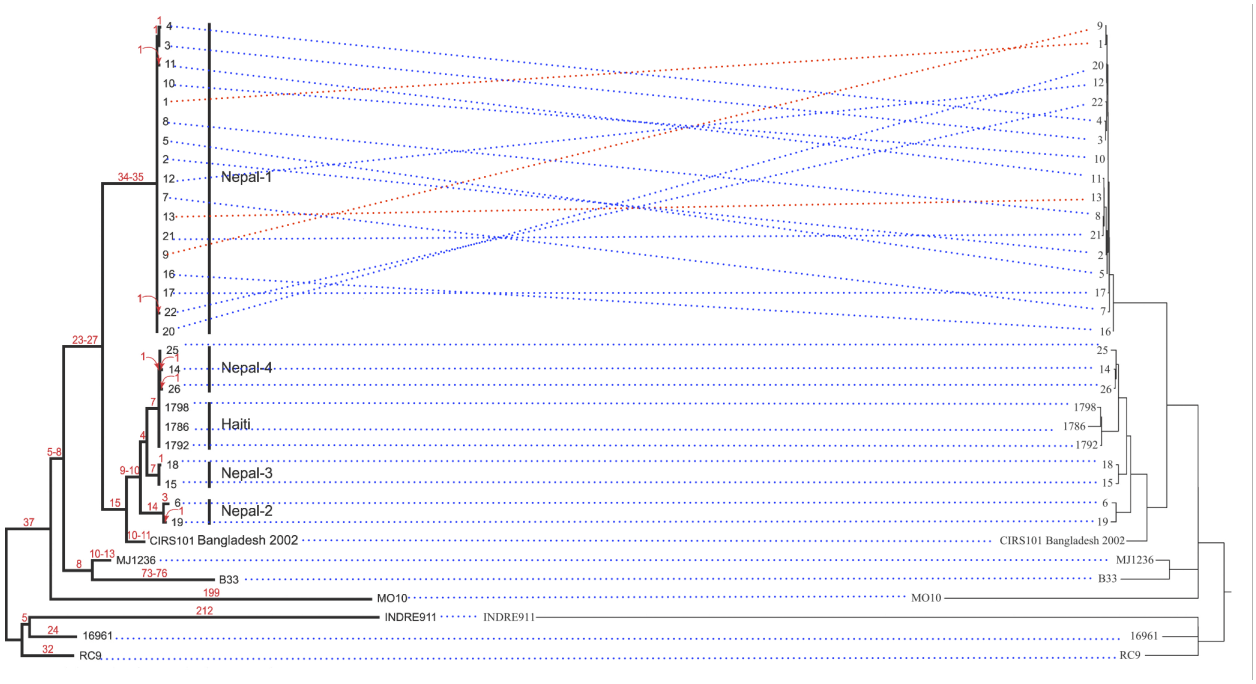


Figure 3 Comparison between phylogenetic trees from published data set (*V. cholerae*) and snpTree server. These trees (34 WGS from ***V. cholerae***) shows comparison of tree topology between the trees from original publication (left) and snpTree server (right). The linked lines indicate exact match for each genome in the tree. According to the tree from published data, the blue lines mean exact match and the red one represent inexact match.

overlapped SNPs). This result supports the consistency of the tree from snpTree server (Figure 3).

S.aureus CC398 data set

For *S. aureus* CC398 (Table 1), snpTree produced a tree with 87 - 88 % concordance for exact match and 96 - 97 % concordance for cluster match. SNP trees for raw reads and assembled genomes are shown in Additional file 3 and Additional file 4 respectively. There were 91 and 90 % overlapping SNPs for raw reads and assembled genomes (Figure 4). The performance of snpTree on this data set was slightly less than for the *V. cholera* data set. The reason is probably that the genomes of 89 *S. aureus* CC398 isolates came from animals and humans sources from 19 countries and four continents. In addition, there are 4,238 SNPs among them [6]. These isolates are more diverse than *V. cholera* isolates. Thus, this diversity makes difficulty for snpTree to capture exactly the same variant as in original publication. Nevertheless, snpTree can differentiate between isolates from humans and pigs which is very meaningful to epidemiological studies.

S. Typhimurium data set

The third data set, *S.* Typhimurium, which consists of 51 *Salmonella* in which 43 isolates from 14 patients with multiple recurrences in Blantyre, Malawi and 8 control typhimurium isolates [11]. Like in the original publication, both raw reads and contigs data set, the isolates fell within three distint phylogenetic clusters (Additional file 5 and 6) which gave 100 % concordance for cluster match (Table 1). On the other hand, the percentage of concordance for exact match was quite low (53 - 61 %). It is not possible to evaluate SNPs position for this data set because of lacking SNPs position data. However, the number of identified SNPs from snpTree server (1,692 SNPs) was not much different from original data set (1,463 SNPs). Most of the *S.* Typhimurium isolates are highly genetically related as they came from patients who had recrudescence and/or reinfections. Therefore, this study requires high-resolution SNPs analysis and intensive phylogenetic tree construction to differentiate these little variation. In addition, the original tree from this data set was generated and confirmed using several independent approaches, with bootstrap support and clade credibility marked [11] which snpTree cannot repeat as using bootstrapping is time-consuming.

M.tuberculosis data set

Another data set that consists of 32 *M. tuberculosis* outbreak isolates and 4 historical isolates (from the same region but isolated before the outbreak) with matching genotype suggesting that the outbreak was clonal [12].

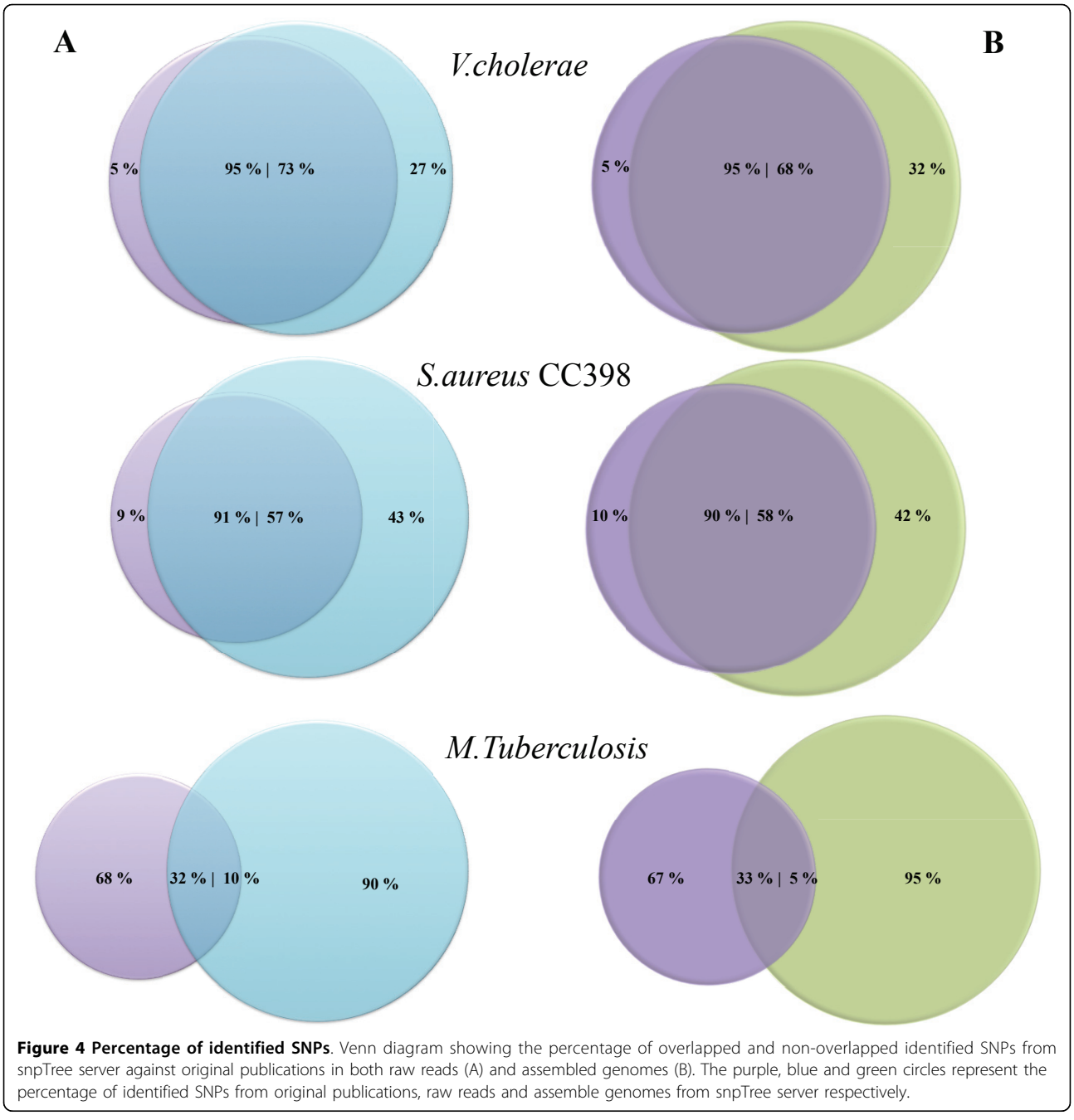


Figure 4 Percentage of identified SNPs. Venn diagram showing the percentage of overlapped and non-overlapped identified SNPs from snpTree server against original publications in both raw reads (A) and assembled genomes (B). The purple, blue and green circles represent the percentage of identified SNPs from original publications, raw reads and assemble genomes from snpTree server respectively.

The performance of snpTree server on this data set was inconsistent due to low concordance percentage for exact match and cluster match (Table 1, Additional file 7 and 8). Moreover, the number of indentified SNPs and matching SNP positions (Figure 3) are very different between the tree from snpTree server (677 SNPs) and the published data (204 SNPs). The original publication determined transmission dynamics of the outbreak at a higher resolution by filtering to remove many of SNPs in repetitive regions and those appearing in a single isolate. Thus, the procedure in the original manuscript is impossible to repeat and it should be noted that the original filtering reduced the number of SNP's from more than 1,000 to 204. This is probably the reason that snpTree were unable to reproduce the same results as in the original publication.

Sensitivity and specificity

In order to evaluate the sensitivity and specificity of SNP calling method, the artificial sequence was created

from a genome of 4,878,012 bp with 1,000 randomly SNP artificial inserted. The simulated sequence was aligned to a reference genome and identified SNPs using SNP idenfication pipeline for assemble genome. SNPs calling was performed with varied two cut-off values which are minimum number of bp between SNPs (prune) and minimum number of bp from a sequence end (e). The sensitivity and specificity for SNP identification were summarized in Table 2.

The sensitivity for prune cut-off (Table 2) was slightly dropped when increasing number of prune. This is due to the more number of bp between SNPs (prune) leading to the high chance to have SNPs between that number of bp.

Using minimum number of bp from a sequence end as a varied cut-off, the sensitivity was very high and stable for all varied values. It is quite rare to have SNPs occurred in the tails of sequence so this cut-off less affects to the SNP calling process. The specificity for both cut-off were very high. It is because the number of SNP inserted is extreamly low (1,000 SNPs) compared to the whole genome (4,878,012 bp).

The rapid technological advantages in WGS and rapidly decreasing cost has made the technology available for large groups of scientists as well as clinical microbiologists. It is expected that WGS will very soon find widespread use in clinical and public health microbiology, as has already been shown [19]. The implementation of such technologies will however, create a major need for simple to use bioinformatic tools to make sense of the data generated. We have here developed snpTree and evaluated it on four different published datasets. The concordance of the SNPs tree from raw reads was more adequate than the one from assembled genomes, which is not surprising. However, in practice transfering sequencing reads will be more time-consuming than just transferring assembled genomes and the tree topology from these different kind of genomes was only sligthly different. Therefore, the assembled genomes option in snpTree server can provide a quicker solution for uploading time-consuming. In order to create informative SNPs tree, using a closely related reference genome is important. Therefore, the selection of a proper reference genome is crucial. Thus, it is adviced to choose a reference genome belonging to the same or as closely related a sub-type as possible to the strain collection under study. This could for species where this is a available reference belonging to the same MLST type. In the future a more generic solution to overcome this obstracle might be to using high-resolution prediction method such as K-mers to assign a genuine reference genome.

Table 2 Sensitivity and specificity

Variable and cut-off value	Sensitivity (%)	Specificity (%)
Number of bp between SNPs		
0	97.8	100
10	97.2	99.99988
25	96.6	99.99975
50	95.8	99.99959
75	94.6	99.99935
100	93.8	99.99918
Number of bp from a sequence end		
0	97.8	100
10	97.8	100
25	97.8	100
50	97.8	100
75	97.8	100
100	97.7	100

Evaluation of sensitivity (SN) and specificity (SP) using different settings of minimum number of bp between SNPs (prune) and minimum number of bp from a sequence end (e) for SNP detection on a simulated dataset consisting of a genome of 4,878,012 bp with 1,000 randomly SNP artificial inserted.

Conclusions

The advance of WGS and the use of epidemiological genomics underline the potential of practical application of WGS for clinical microbiology and emphazies the importance of biology and evolution in developing reliable and accurate genomics tools for clinical use. In addition, SNP-typing phylogenetic methods can distinguish very closely related isolates to a degree not achievable by widely employed sub-genomic typing tools. snpTree server might be not a perfect tool but it is an option for easy and rapid standardised and automatic SNP analysis tool in epidemiological studies. It is also useful for users with limited bioinformatic experience.

Additional material

Additional file 1: Example of SNP annotation output.

Additional file 2: SNP trees from contigs of *V. cholerae* data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 3: SNP trees from raw reads of *S. aureus* CC398 data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 4: SNP trees from contigs of *S. aureus* CC398 data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 5: SNP trees from raw reads of *S.* Typhimurium data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 6: SNP trees from contigs of *S.* Typhimurium data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 7: SNP trees from raw reads of *M. tuberculosis* data set (left is the tree from original publication and right is the tree from snpTree server).

Additional file 8: SNP trees from contigs of *M. tuberculosis* data set (left is the tree from original publication and right is the tree from snpTree server).

Acknowledgements
This study was supported by the Center for Genomic Epidemiology (09-067103/DSF) http://www.genomicepidemiology.org and Danish Food Industry Agency (3304-FVFP-08). PL and RKM would like to acknowledge funding from the Technical University of Denmark.
This article has been published as part of *BMC Genomics* Volume 13 Supplement 7, 2012: Eleventh International Conference on Bioinformatics (InCoB2012): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S7.

Author details
[1]National Food Institute, Building 204, Technical University of Denmark, 2800 Kgs Lyngby, Denmark 4444. [2]Center for Biological Sequence Analysis, Building 208, Department of Systems Biology, Technical University of Denmark, 2800 Kgs Lyngby, Denmark.

Authors' contributions
PL planned the study, carried out web-server construction and drafted the manuscript. RKM constructed SNPs analysis pipeline for assembled genomes and automatic SNP tree construction pipeline. MCFT participated in web-server construction. CF constructed automatic SNPs tree construction pipeline. SR constructed SNPs analysis pipeline for raw reads and developed Genobox toolbox. FMA supervised, planned the study and drafted the manuscript. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 13 December 2012

References

1. Parkhill J, Wren BW: **Bacterial epidemiology and biology–lessons from genome sequencing.** *Genome biology* 2011, **12**:230.
2. Foxman B, Zhang L, Koopman JS, Manning SD, Marrs CF: **Choosing an appropriate bacterial typing technique for epidemiologic studies.** *Epidemiologic perspectives & innovations* 2005, **2**:10.
3. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM: **Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak.** *MBio* 2011, **2**.
4. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327**:469-74.
5. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP: **Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:3065-70.
6. Price LB, Stegger M, Hasman H, Aziz M, Larsen J, Andersen PS, Pearson T, Waters AE, Foster JT, Schupp J, Gillece J, Driebe E, Liu CM, Springer B, Zdovc I, Battisti A, Franco A, Zmudzki J, Schwarz S, Butaye P, Jouy E, Pomba C, Porrero MC, Ruimy R, Smith TC, Robinson DA, Weese JS, Arriola CS, Yu F, Laurent F, Keim P, Skov R AF: **Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock.** *MBio* 2012, **3**:1-6.
7. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J: **SNP detection for massively parallel whole-genome resequencing.** *Genome Res* 2009, **19**:1124-1132.
8. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome research* 2010, **20**:1297-303.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-9.
10. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic acids research* 2002, **30**:2478-83.
11. Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, Parkhill J, Dougan G, Gordon MA: **High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal salmonella typhimurium disease.** *Clinical infectious diseases* 2012, **54**:955-63.
12. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *The New England journal of medicine* 2011, **364**:730-9.
13. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-60.
14. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O: **Multilocus Sequence Typing of Total Genome Sequenced Bacteria.** *Journal of clinical microbiology* 2012, 1355-1361.
15. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nature reviews. Genetics* 2011, **12**:443-51.
16. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.** *Molecular biology and evolution* 2009, **26**:1641-50.
17. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**:2156-8.
18. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome research* 2008, **18**:821-9.
19. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW: **A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance.** *BMJ Open* 2012, **2**.

doi:10.1186/1471-2164-13-S7-S6
Cite this article as: Leekitcharoenphon *et al.*: snpTree - a web-server to identify and construct SNP trees from whole genome sequence data. *BMC Genomics* 2012 **13**(Suppl 7):S6.

Article IV

Genomic epidemiology of the global occurrence of *S.* Typhimurium DT104.

ARTICLE IV

Pimlapas Leekitcharoenphon,[1,2*] Rene S. Hendriksen,[1] Ole Lund,[2] Frank M. Aarestrup,[1]

[1] Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs Lyngby, Denmark

*Corresponding author: **Pimlapas Leekitcharoenphon,**
Division for Epidemiology and Microbial Genomics,
National Food Institute, Technical University of Denmark,
Kgs. Lyngby, Denmark
E-mail: pile@food.dtu.dk

Manuscript

Genomic epidemiology of the global occurrence of *S.* Typhimurium DT104

Pimlapas Leekitcharoenphon[1,2], Rene S. Hendriksen[1], Ole Lund[2], and Frank M. Aarestrup[1]

Abstract— **It has been thirty years since the initial emerging and subsequent rapid global spread of multidrug-resistant *S.* Typhimurium DT104. Nonetheless, its origin and transmission route have never been revealed. We used whole genome sequence (WGS) and temporally structured sequence analysis within Bayesian framework to reconstruct temporal and spatial phylogenetic trees, estimate rate of mutation and divergence time of 315 *S.* Typhimurium DT104 isolates sampled from 1969 to 2012 from twenty-one countries in six continents. The DT104 was estimated to initially emerge as antimicrobial-susceptible strains in ~1946 (1931 - 1959) and further became multidrug-resistant (MDR) DT104 in ~1974 (1966 – 1981) through horizontal transfer of the 13-kb SGI1 MDR region into already present SGI1-contained susceptible strains. This was followed by multiple transmission events initially from Central Europe and later between European countries. An independent transmission occurred to USA and another to Japan and from here to Taiwan and Canada. An independent acquisition of resistance took place in Thailand in ~1986 (1975 – 1990). Our study also confirms that DT104 most likely spreads among food animals and from here transmit to humans and they do not exhibit different epidemics. Locally in Denmark, WGS was capable to confirm local epidemiology for transmission between animal herds. Interestingly, the demographic history of Danish MDR DT104 provided evidence for the accomplishment of an eradicating program across pig herds in Denmark in 1996 to 2000. The results from this study would suggest any potential monitor and strategies for further prevention and control of similar successful clones.**

INTRODUCTION

Salmonella is one of the most common foodborne pathogens worldwide [1]. In the US alone, salmonellosis was estimated to cause 1.4 million cases effecting 17,000 hospitalizations and almost 600 deaths each year [2,3]. Globally, *Salmonella enterica* serovar Typhimurium is the most commonly isolated serovar [1]. *S.* Typhimurium consists of a number of subtypes that classically have been divided by phagetyping. During the last three decades, *S.* Typhimurium phage type DT104 emerged as the most important phage type and one of the best-studied because of its rapid global dissemination [1,4]. One of the characteristics of DT104 is its typically resistance to ampicillin, chloramphenicol, streptomycin, sulfonamide, and tetracycline (ACSSuT)[5] and its capacity to acquire extra resistance to other clinically important antimicrobial drugs[4].

Susceptible DT104 was first reported in 1960s, and subsequently as multidrug-resistant (MDR) DT104 in the early 1980s in the United Kingdom from humans and birds [6,7 8,9]. The first report on isolates from agricultural animals were in the UK in 1988 [8] and in the US in 1990 [10]. MDR DT104 rapidly emerged globally in 1990s and became the most prevalent reported phage type from humans and animals in many countries [4,6,11]. Previous epidemics with MDR phage types of *S.* Typhimurium, such as DTs 29, 204, 193 and 204c, were mostly restricted to cattle, whereas MDR DT104 spread among all domestic animals including cattle, poultry, pigs and sheep [6]. A decline in MDR DT104 has been reported in the last decade [12,13].

A recent study used WGS to study DT104 from mainly cattle and humans in Scotland[14]. This study was severely hampered by the lack of inclusion of isolates from other animal species and by not including the fact that infections in humans are from food products of which most consumed in Scotland are imported from other countries[14,15].

Despite several studies show that the origin and transmission routes of the phage type DT104 are still ambiguous. Based on the presence of the rare resistance genes floR and tet(G), it has been suggested that the MDR phage type originated in South East Asia[6]. The transmission has been suggested to be through trade with live animals, but it has never been established whether the epidemiology in the different animal species are part of a common global spread or whether there are host specific variants.

In order to get closer answers to these questions, we sequenced a carefully selected representative intercontinental DT104 collection from different sources in twenty-one countries covering the period from 1969 to 2012. We identified SNPs and phylogenomic dating based on temporally structured sequence analysis within a Bayesian framework aiming to exhibit population structure, phylogeny and evolution over time of DT104. Besides, we also revealed historically as well as very recent disseminations events and locally between and within farms in Denmark.

RESULTS

A global collection of 315 *S.* Typhimurium DT104 isolates was sampled from 1969 to 2012. The collection represented Europe (n=235), Asia (n=48), Australia (n=7), North America (n=18), South America (n=5) and Africa (n=2). The isolates

[1]Division for Epidemiology and Microbial Genomics, National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark.

[2]Center for Biological Sequence Analysis, Department of System Biology, Technical University of Denmark, Kgs. Lyngby, Denmark.

were from animal (n=196) and human sources (n=119). Seventy-five of the animal isolates were from Denmark and selected to cover animal hosts, temporal and spatial diversity as well as specific epidemiological events that had been left unexplained during the last 20 years investigation of DT104 in Denmark. The complete information of the studied isolates can be found in Supplementary Table 1.

Global phylogeny of S. Typhimurium DT104

A global collection of DT104 isolates was subjected to WGS and 4,619 qualified SNPs were identified. We identified no significant recombination in this collection. We applied phylogenomic dating to reconstruct temporal and spatial phylogenetic tree using BEAST (Bayesian Evolutionary Analysis Sampling Trees) [16,17]. A combination of Bayesian Skyline model and relaxed uncorrelated lognormal clock were selected as population size change and molecular clock models. Bayesian based tree for all 315 DT104 isolates is showed in Figure 1a. The mutation rate was estimated to be 2.97 x 10^{-7} SNP/site/year that was approximated to 1.47 SNP/year. The estimated rate of mutation corresponds to the mutation rates from previous studies of invasive *S.* Typhimurium in sub-Saharan Africa [18] and multidrug-resistant *S.* Typhimurium DT104 in different hosts [14]. The most recent common ancestor was estimated to emerge in 1946 (95% highest posterior density, HPD, 1931 - 1959). The tree consisted of two individual clusters; a cluster of susceptible and resistant isolates and a complex cluster of multidrug-resistant strains with resistance to ampicillin, chloramphenicol, streptomycin, sulfonamide and tetracycline (ACSSuT resistance type). The susceptible and MDR clusters differed approximately by 109 SNPs. An average SNP difference among isolates in the susceptible cluster (n=18) was 103 SNPs, whereas that number among isolates in MDR cluster was only 60 SNPs (38 – 100 SNPs) despite a large number of isolates in the MDR cluster (n=297). In contrast to the MDR strains, all of the isolates in the susceptible cluster contained small fragment or partial sequences of the 43-kb *Salmonella* genomic island 1 (SGI1, GenBank accession number AF261825) [19,20] and none of them harbored the 13-kb SGI1 multidrug resistance region [21].

By using comparative genomics, we found 4,472 core genes from the DT104 collection meaning that 96% of total genes in a genome are common among DT104 strains. This number of core genes is relatively higher than the number of 62% of genes found commonly within *Salmonella enterica* [22]. Core gene sequences can be obtained from Supplementary 2.

Based on the temporal phylogenetic tree, the proposed transmissions were illustrated in Figure 2. From an unidentified source, *S.* Typhimurium DT104 originated as a susceptible strain in 1946. Susceptible strains later emerged in Morocco, Spain and France in ~1959 (95% HPD 1956-1968). In ~1971 (95% HPD 1957-1977), the unknown source-susceptible DT104 appeared in Thailand where it likely was further transferred to Denmark in ~1996 (95% HPD 1988-2002). Locally in Thailand the susceptible strains evolved as resistant in ~1986 (95% HPD 1975-1990).

The 261 MDR isolates were analyzed separately yielding 3,621 variable sites for Bayesian tree construction using BEAST (Figure 1b). The European isolates disseminated throughout the tree whereas the isolates from the other continents seem to be restricted to their continental origins except the human isolates from New Zealand that spread throughout the tree and clustered with isolates from different countries and continents (Figure 1b) suggesting that they might be travel-related cases. This result is concordant with the report that Australia and New Zealand have had few MDR DT104 human infections and most of human cases were from travellers [4]. Another study found that 37% of Australian DT104 isolates were associated with travel aboard, especially to Southeast Asia [4].

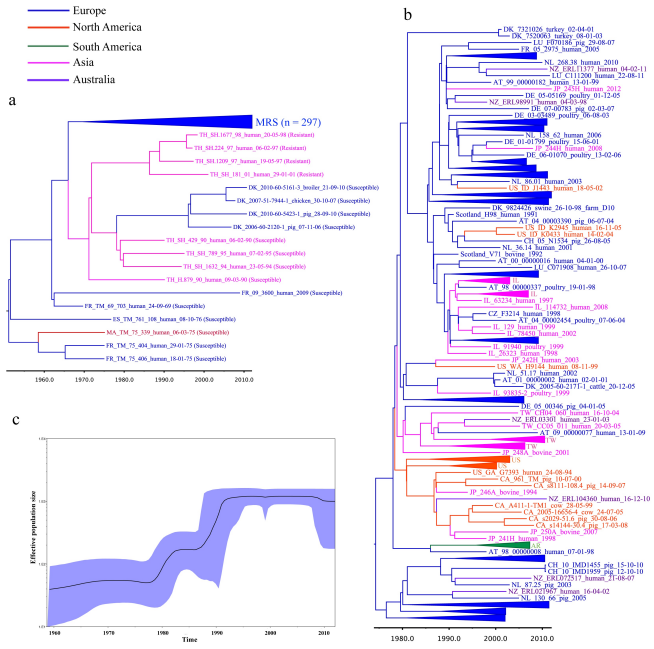


Figure 1 Global phylogeny of *S.* Typhimurium DT104. Bayesian based temporal phylogenetic trees from BEAST of (**a**) all DT104 and (**b**) sub-sampled MDR DT104 isolates. The tree in (**a**) showed the most recent common ancestor of *S.* Typhimurium DT104 in ~1946 (1931 - 1959) and exhibited distinct clusters between a susceptible DT104 cluster and MDR DT104 cluster. Meanwhile, the tree in (**b**) estimated that MDR DT104 initially emerged in ~1974 (1966 – 1981). The changes in effective population size over time (year) were illustrated in Bayesian skyline plot (**c**). Isolates were named by country of origin, isolate ID, source, and date (dd-mm-yy). Branches and nodes were colored according to the continent of isolate. Country abbreviations were used as follow. AR; Argentina, AT; Austria, CA; Canada, CZ; Czech Republic, DK; Denmark, FR; France, DE; Germany, IE; Ireland, IL; Israel, JP; Japan, LU; Luxemburg, MA; Morocco, NL; The Netherlands, NZ; New Zealand, PL; Poland, ES; Spain, CH; Switzerland, TW; Taiwan; TH; Thailand, US; The United States.

MDR DT104 was estimated to appear in ~1974 (95% HPD 1966 – 1981) (Figure 1b and Figure 2). From an unknown-source multiple introductions of MDR DT104 occurred to

Europe from ~1976 (95% HPD 1975-1984). Subsequently another introduction to and from Israel occurred in ~1990 (95% HPD 1987-1994). Separated transmission routes occurred to Japan in ~1980 (95% HPD 1977-1985) and from Japan to Taiwan in ~1983 (95% HPD 1981-1988) and from Japan to Canada in ~1987 (95% HPD 1986-1991). In addition, the tree suggested that unknown-source MDR DT104 initially spread to the United States in ~1980 (95% HPD 1978-1987), consistent with the report of the emergence of MDR DT104 in the United States, particular in western states in early 1985 [23]. Furthermore, it spread from Israel to Argentina in ~1984 (95% HPD 1976-1990) with 81 average SNP difference (Figure 2).

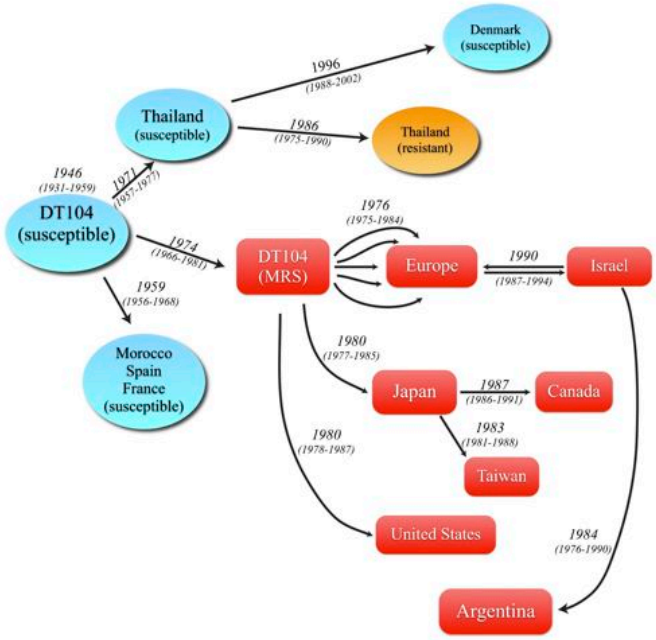


Figure 2 Diagram of the dissemination of *S*. Typhimurium DT104. Ages of nodes and divergence time of interested events from Figure 1a and 1b were summarized and illustrated in this diagram. The unknown-source *S*. Typhimurium DT104 initially emerged as susceptible strains in ~1946 (1931 – 1959). The susceptible DT104 was estimated to become MDR DT104 in ~1974 (1966 – 1981). The MDR DT104 from unknown source spread to Europe and other continents in ~1976 and 1980s respectively. Estimated time when transmission initially occurred (year) are represented as the median values, with 95% HPD in parenthesis.

Bayesian skyline plot for all DT104 isolates showed a demographic history of DT104 from ~1960 (Figure 1c). The effective population size of DT104 rose gradually until ~1980 after it became MDR DT104, and the population size increased sharply from 1980 to 1985 (Figure 1c). This coincides with the estimated time of the occurrence of MDR DT104 in ~1974 (Figure 2) and the initial dissemination of MDR DT104 throughout Europe, Asia and America during 1980s (Figure 2). The second wave of DT104 started in ~1990, and the population size increased dramatically. This increasing may reflect the global dissemination of MDR DT104 because the timeline is agreeable with the occurrences of MDR DT104 in many countries. Germany had an increase in DT104 in the beginning of 1990s [24,25]. The number of DT104 human infections in UK rose from 259 in 1990 to 4006 in 1995 [26] as well as the number of DT104 in animals increased from 458 in 1993 to 1513 in 1996 [7]. Almost all 67% of *Salmonella* isolates from animals in Scotland during 1994-1995 were MDR DT104 [27], and a number of studies showed that throughout the 1990s, MDR DT104 spread to other parts of the world, including the United States, the United Kingdom, and France [23,28–30].The trend has leveled off since 1995 and gradually decreased from 2008.

Dissemination of DT104 in Europe

The spatial and temporal transmissions of the animal MDR DT104 isolates within European countries are summarized and illustrated in Figure 3. The earliest predicted disseminations (Figure 3a) were from Germany to Czech Republic in ~1984 (95% HPD 1982-1988), from Germany to Denmark in ~1985 (95% HPD 1982-1990) and from Germany to Scotland in ~1986 (95% HPD 1984-1989). More recent disseminations were from Denmark backward to Germany in ~1988 (95% HPD 1987-1994) and Germany to Netherlands in ~1988 (95% HPD 1984-1990). In addition, Germany had outward phylogenetic link to Israel in ~1988 (95% HPD 1986-1991) because of isolates from poultry. The next waves (Figure 3b) were from Netherlands to Ireland in ~1992 (95% HPD 1988-1997) and Switzerland in ~1993 (95% HPD 1988-1997). In early 1990s, Denmark had outward phylogenetic links to Poland in ~1992 (95% HPD 1988-1996), Austria in ~1992 (95% HPD 1990-2000), Luxemburg in ~1993 (95% HPD 1988-1997) and Ireland in ~1993 (95% HPD 1989-2001). In the same period, Germany had outward links to Luxemburg in ~1990 (95% HPD 1990-1998), Austria in ~1990 (95% HPD 1988-1996) and Switzerland in ~1993 (95% HPD 1990-1997). Another hub in early 1990s was Scotland where the potential disseminations linked to Austria in ~1990 (95% HPD 1987-1991), Ireland in ~1990 (95% HPD 1986-1994), Netherlands in ~1991 (95% HPD 1989-1993), Denmark in ~1992 (95% HPD 1988-1994) and Switzerland in ~1993 (95% HPD 1989-1995). Scotland is a net importer of food [14] for instance 58% of all red meat and 38% of raw beef are non-Scottish origin [15]. Austria also had phylogenetic link back to Denmark in ~1998 (95% HPD 1990-1999) and had an achievable link to Israel in 1992 (95% HPD 1989-1994) via isolates from poultry. The most recent predicted transmission was from Scotland to Luxemburg in ~2000 (95% HPD 1998-2005).

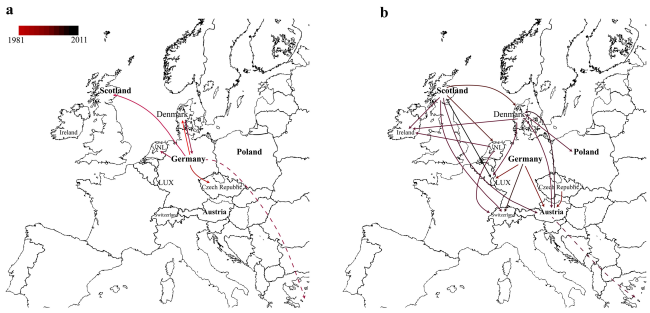


Figure 3 Transmission within Europe of MDR *S.* Typhimurium DT104 from animal isolates. Discrete phylogeographic analysis of MDR DT104 during 1981 to 1990 (**a**) and 1990 to 2011 (**b**) within European countries. Locations and transmission lines were obtained from nodes and branches in BEAST analysis. The color gradient is represented the ages of transmission lines.

Local phylogeny of S. Typhimurium DT104

Seventy-five MDR *S.* Typhimurium DT104 isolates sampled from 1997 to 2011, originating from several farms in Denmark were sequenced. Sequence alignments of 755 SNPs were analyzed using BEAST. The Bayesian phylogenetic tree (Figure 4a) established an estimated mutation rate at 2.15 x 10^{-7} SNP/site/year or 1.06 SNPs per year. The most recent common ancestor was predicted to emerge at the same period with the occurrence of the global MDR DT104 in ~1974 (95% HPD 1966 – 1981). The tree was divided into two major clusters and subsequently branched off to many lineages indicating multiple introductions of MDR DT104 to different farms in Denmark.

Several isolates were selected from the same farms. Most of those isolates were clustered phylogenetically according to their farms. Isolates from four different farms namely D32, D41, D42 and D47 were mixed into the same lineage. This is consistent with the information that there has been physical contact among those four farms, thus showing the ability of WGS to confirm very local epidemiology. There were several branching links between isolates from swine and cattle (Figure 4a), whereas isolates from poultry clustered separately. This indicates free transmission between cattle and swine, but a more closed spread in the poultry production. Concordantly, the analysis of proliferation of the infection in various species suggested that DT104 strains spread from cattle to pigs and humans [7,31].

The relation between population structure and time (Figure 4b) showed that the effective population size of MDR DT104 in Denmark rose slowly until ~1984 then it increased sharply from ~1984 to ~1987. Subsequently, the population was firmly established until ~1998 and it declined dramatically during ~1999 to ~2000, when an intensive eradication program was attempted in Denmark [32]. Following the abandon of the eradication program, the population size increased in ~2001 and decreased slightly from ~2004. We carried out different Bayesian skyline plots based on sources (Supplementary 3). The pattern of sharp decline during 2000 has not been found among isolates from cattle, poultry and human except isolates from swine. In fact, 69% of Danish isolates were swine. Thus, the decline of the population size in 2000 was related to swine isolates.

Discrete phylogeographic analysis indicated several relationships among farms in Denmark. The complete phylogeographic link can be found in Supplementary 4. Average SNP distance between farms ranged from 3 to 100 SNPs. We have four confirmed physical contacts between farms. Those contacts were concordant to the phylogeographic links showed in Figure 4c. The contacts between farms D12-D38 and D41-D42 were direct relationships with 30 and 7 SNPs differences respectively, whereas the contacts from farms D32-D42 and D42-D47 were indirect contacts employed by 10 and 8 SNPs distances respectively. Interestingly, data from one farm (D10) where eradication was presumed unsuccessfully performed showed that isolates found after eradication was not the same lineage as the isolates prior to eradication.

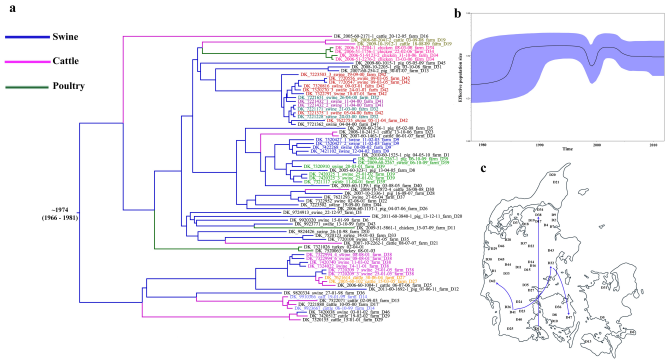


Figure 4 Local phylogeny of MDR *S.* Typhimurium DT104 isolates in Denmark. Bayesian based phylogenetic tree of 75 Danish MDR DT104 (a) showed that the most recent common ancestor was estimated to emerge in ~1974 (95% HPD 1966 – 1981). The tree was further divided into two major clusters in ~1973 (HPD 1961-1983) and ~1976 (HPD 1964-1985). Farm numbers were noted at the end of node names. Nodes were colored according to farm of origin. A single isolate from a single farm was labeled in black. Colored branches showed animal sources. Bayesian skyline plot of changes in population size of Danish MDR DT104 over time was showed in (b). Geographic diffusion across different farms based on discrete phylogeographic analysis for the confirmed-farm contacts was illustrated in (c). The complete geospatial transmission can be retrieved from Supplementary 4.

DISCUSSIONS

Global epidemiology

S. Typhimurium DT104 has gained intensive global interest due to its rapid intercontinental dissemination, the chromosomal location of multiple resistance genes and its capacity to promptly acquire additional resistance traits [4].

Our analysis of a global collection of DT104 suggest that the most recent common ancestor of *S.* Typhimurium DT104 emerged in ~1946 (1931 - 1959) as antimicrobial-susceptible DT104 (Figure 1A) in an unidentified reservoir. The earliest

reports on susceptible DT104 strains isolated from human infections appeared in 1960s in the United Kingdom [6]. However, most if not all non-typhoidal *Salmonella* serovars have their natural reservoir in animals and only occasionally infect humans. Thus, susceptible DT104 may easily have spread for several years in an animal reservoir before the first infections occurred in humans. Interestingly, our results suggest that the unknown source-susceptible DT104 spread to Thailand in ~1971 (1957-1977) and later locally acquired resistance in ~1986 (1975-1990) in Thailand (Figure 1A and 2). It has previously been assumed that these resistant isolates have emerged from MDR strain that have lost some of the resistance genes. However, this study contradicts this hypothesis.

Our result suggests that DT104 initially became multidrug-resistant DT104 in ~1974 (1966 – 1981) from unknown source (Figure 1B). The first observations of MDR DT104 were in seagull and cattle in the UK in 1984 [6,24,33], where it was thought to have originated from gulls and exotic birds imported from Indonesia and Hong Kong [6]. An Asian origin have also been suggested in other previous studies, where it have been indicated that the resistance determinants of MDR DT104 strains may have emerged among bacteria in aquaculture and subsequently been horizontally transferred to *S*. Typhimurium DT104 [34]. Since most of farmed shrimp are produced in Asia, in particularly China and Thailand, it was suggested that the emergence of Thai resistant DT104 might be caused by aquaculture bacteria. Our study contradicts this hypothesis. Based on our results, a European origin of MDR DT104 seems much more likely. Thus, the isolates from Thailand are not involved in the MDR DT104 cluster and MDR DT104 did not emerge in the countries from which we have isolates prior to 1980.

The Bayesian phylogenetic tree revealed that the susceptible and MDR clusters differed by 109 SNPs indicating that these two clusters are diverse. The 18 isolates within susceptible cluster had 103 SNP differences while there were 60 SNP distances within MDR cluster (n=297) suggesting that the MDR strains have higher degree of clonality. From sequence comparison, we found that the susceptible strains contained a partial 43-kb *Salmonella* genomic island 1 (SGI1) and none of them harbored a 13-kb SGI1 MDR region. One of the SGI1 functions is an integrative mobilizable element [35] and the DT104 drug resistance genes can be transduced by P22-like phage ES18 and by phage PDT17, which are produced so far by all DT104 isolates [36]. The emergence of MDR strains would therefore cause by horizontal transfer of the DT104 antibiotic resistance gene cluster [37] into the SGI1-contained susceptible strains. The good evidence for horizontal transfer of the antibiotic resistance gene cluster is the presence of this cluster in another *S. enterica* serovar Agona [38]. Our result challenges the hypothesis that the MDR DT104 emerged by acquiring an entire SGI1 with MDR region[37].

Local epidemiology

The temporal phylogenetic tree (Figure 4a) estimated that the most recent common ancestor of Danish MDR DT104 initially emerged in ~1974 (1966 – 1981). This emerging time was before the earliest emerge of MDR DT104 in the United Kingdom in early 1980s indicating that MDR DT104 had been in Denmark for several years without causing tremendous spread of infection as well as the observation in 1998 found that the Danish isolated of MDR DT104 recovered from 1991 to 1995 were very similar to those found from 1996 to 1998 [32].

The Bayesian phylogenetic tree showed the capacity to cluster isolates from the same herd and to cluster isolates from different confirmed contact farms suggesting that WGS is useful for locally epidemiological observation across animal herds.

Changes in effective population size over time provided an interesting point that there was a sharp decline of the population size of swine isolated MDR DT104 during ~1999 to ~2000 and a sharp increase of the population size to the same state prior decreasing since ~2001. The decreasing of swine MDR DT104 is an evidence of the accomplishment of the eradicating program in 1996 to 2000 established by the Federation of Danish Pig Producers and Slaughterhouse, in collaboration with the Danish Veterinary Service and the Danish Veterinary Laboratory. The program aimed to eradicate MDR DT104 from infected pig herds. The methods used included the depopulation of pig herds and the cleaning and disinfection of building before repopulation with pig free from DT104 [32].

CONCLUSIONS

This study shows the timeline of global and local disseminations of *S*. Typhimurium DT104 and the evolution of antimicrobial susceptible strains to MDR DT104 strains through horizontal transfer of 13-kb SGI MDR region. The results are consistent with many historical occurrences of MDR DT104 since it was observed in 1984. Moreover, the results carried out by WGS also confirm local epidemiology of DT104 and the efficiency of eradicating program in Denmark. The predicted transmission routes and demographic history would suggest any potential monitor and strategies for further prevention and control of similar successful clones.

METHODS

Bacterial isolates

A total of 315 *S*. Typhimurium DT104 isolates included in this study were received intercontinentally from 21 countries; Argentina (n=5), Austria (n=30), Canada (n=6), Czech Republic (n=9), Denmark (n=79), France (n=9), Germany (n=27), Ireland (n=10), Israel (n=17), Japan (n=10), Luxemburg (n=13), Morocco (n=2), The Netherlands (n=22), New Zealand (n=7), Poland (n=13), Scotland (n=14), Spain (n=1), Switzerland (n=8), Taiwan (n=13), Thailand (n=8) and The United States (n=12). All isolates from Japan and Scotland were retrieved as paired-end reads from the recent study [14] via European Nucleotide Archive. The rest of isolates were supplied from the laboratory strain collections in the

respective countries. The time spanning of the isolates ranged from 1969 to 2012, which the most antique isolates were human isolate from France in 1969, human isolates from Morocco in 1975 and 1981 and human isolate from Spain in 1976. Isolates were from various sources; cattle (n=35), poultry (n=51), swine (n=109), hare (n=1) and humans (n=119). The full information of isolates is shown in Supplementary Table 1.

Whole genome sequencing, de novo assembly and resistance genes

Isolates were either sequenced using Illumina HiSeq or MiSeq. Raw sequence data have been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession no. xxxxxx. The raw reads were *de novo* assembled using the pipeline available on the Center for Genomic Epidemiology (CGE) (www.genomicepidemiology.org), which is based on Velvet algorithms for *de novo* short reads assembly [39]. A complete list of genomic sequence data is available in the Supplementary Table 1. The assembled genomes were further analyzed using similar pipeline available on the CGE website. The web-servers ResFinder [40] were used to detect acquired antimicrobial resistance genes with a selected threshold equal to 80 % identity.

SNP identification

Single nucleotide polymorphisms (SNPs) were determined using a genobox pipeline available on the Center for Genomic Epidemiology (www.genomicepidemiology.org) [41]. Fundamentally, the pipeline consists of various publicly available programs. The paired-end reads were aligned against the reference genome, *S.* Typhimurium DT104 (accession number HF937208, genome length 4,933,631 bp) [14], using Burrows-Wheeler Aligner (BWA) [42]. SAMtools [43] 'mpileup' commands were used to identify and filter SNPs. The qualified SNPs were selected once they met the following criteria: (1) a minimum coverage (number of reads mapped to reference positions) of 5; (2) a minimum distance of 15 bps between each SNP; (3) a minimum quality score for each SNP at 20; and (4) all indels were excluded. The final qualified SNPs for each genome were concatenated to an alignment by an in-house python script.

Temporal Bayesian phylogeny, discrete phylogeographic analysis and Bayesian skyline plot

SNP alignments were subjected to Bayesian Evolutionary Analysis Sampling Trees, BEAST version 1.7 [16,17] for temporal phylogenetic reconstruction, estimation of mutation rate and divergence time. Several combinations of population size change and molecular clock models were evaluated to find the best-fit models. Among tested models, the combination of a skyline model [44] of population size change and a relaxed uncorrelated lognormal clock gave the highest Bayes factors. The selected models allow the evolutionary rates to change [45] among the branches of the tree, and a GTR substitution model with γ correction for among-site rate variation.

All BEAST simulations were run for at least 150 million and up to 300 million steps, subsampling every 10,000 steps. The trees produced by BEAST were summarized onto a single target tree using TreeAnnotator [17] with 10% of the MCMC chains discarded as burn-in. Statistical confidence is represented by values for the 95% highest probability density (HPD). A final tree was viewed and edited in FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The geographic locations and direction of the transmissions were estimated by the discrete phylogeographic analysis using a standard continuous-time Markov chain (CTMC) [46] implemented in BEAST. A location-annotated MCC tree was converted to KML format using phylogeo.jar, which is relatively equivalent to SPREAD (http://www.phylogeography.org/SPREAD.html). The KML file was visualized in Google Earth (http://earth.google.com/).

Demographic history was reconstructed using Bayesian skyline plot implemented in Tracer [17] by estimating the genealogy and inferring the effective population size at different points along the genealogy timescale. The population size was inferred by the product of the interval size (γ_i) and $i(i - 1)/2$, where i is the number of genealogical lineages in the interval [47,48].

ACKNOWLEDGES

This study was supported by the Center for Genomic Epidemiology (09-067103/DSF) http://www.genomicepidemiology.org. The authors would like to acknowledge sixteen institutes to provide the DT104 isolates used in this study; (1) Servicio Enterobacterias, Departamento Bacteriología, INEI - ANLIS "Dr. Carlos G. Malbrán", Buenos Aires, Argentina. (2) Institute Austrian Agency for Health and Food Safety (AGES) NRC Salmonella, Austria. (3) Public Health Agency of Canada, Laboratory for Foodborne Zoonoses, Canada. (4) Czech Republic. (5) Federal Institute for Risk Assessment (BfR), Department of Biological Safety, National Reference Laboratory for Salmonella (NRL-Salm), Germany. (6) National Reference Laboratory Salmonella, Department of Agriculture, Food and the Marine Laboratories, Kildare, Ireland. (7) Government Central Laboratories, Jerusalem, Israel. (8) Surveillance Epidémiologique, Laboratoire National de Santé, Luxemburg. (9) Central Veterinary Institute (CVI) part of Wageningen UR, Lelystad, The Netherlands. (10) Enteric Reference Laboratory and *Leptospira* Reference Laboratory, ESR (Institute of Environmental Science and Research Ltd), New Zealand. (11) Department of Microbiology, National Reference Laboratory for Salmonellosis, National Veterinary Research Institute, Poland. (12) Institute of veterinary bacteriology, the Centre for Zoonoses, Bacterial Animal Diseases and Antimicrobial Resistance (ZOBA), Berne, Switzerland. (13) Centers for Disease Control, Taiwan. (14) Department of Medical Sciences, WHO International Salmonella and Shigella Centre, National Institute of Health, Ministry of Public Health,

Bangkok, Thailand. (15) PulseNet Next Generation Subtyping Methods Unit, NCEZID/DFWED/EDLB, Centers for Disease Control and Prevention, Atlanta, GA, The United States. (16) Center for Veterinary Medicine, US Food and Drug Administration, Laurel, Maryland, The United States. In addition, the authors would like to thank Jessica Hedge for advices on BEAST program.

REFERENCES

1. Lan, R., Reeves, P. R. & Octavia, S. Population structure, origins and evolution of major Salmonella enterica clones. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **9**, 996–1005 (2009).
2. Voetsch, A. C. *et al.* FoodNet estimate of the burden of illness caused by nontyphoidal Salmonella infections in the United States. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **38 Suppl 3**, S127–34 (2004).
3. Hendriksen, S. W. M., Orsel, K., Wagenaar, J. A. & Miko, A. Salmonella Typhimurium DT104A Variant. **10**, 2225–2227 (2004).
4. Helms, M., Ethelberg, S. & Mølbak, K. International Salmonella Typhimurium DT104 infections, 1992-2001. *Emerging infectious diseases* **11**, 859–67 (2005).
5. Mulvey, M. R., Boyd, D. A., Olson, A. B., Doublet, B. & Cloeckaert, A. The genetics of Salmonella genomic island 1. *Microbes and infection / Institut Pasteur* **8**, 1915–22 (2006).
6. Threlfall, E. J. Epidemic Salmonella typhimurium DT 104 — a truly international. *J Antimicrob Chemother* **46(1)**, 7–10 (2000).
7. Poppe, C. *et al.* Salmonella typhimurium DT104: a virulent and drug-resistant pathogen. *The Canadian veterinary journal. La revue vétérinaire canadienne* **39**, 559–65 (1998).
8. Threlfall, E. J., Frost, J. A., Ward, L. R. & Rowe, B. Epidemic in cattle and humans of Salmonella typhimurium DT 104 with chromosomally integrated multiple drug resistance. *The Veterinary record* **134**, 577 (1994).
9. Hollinger, K. *et al.* Salmonella Typhimurium DT104 in cattle in Great Britain. *Journal of the American Veterinary Medical Association* **213**, 1732–3 (1998).
10.Hancock, D. Besser, T. Gay, J. Rice, D. Davis, M. Gay, C. The global epidemiology of multiresistant Salmonella enterica sero- var Typhimurium DT104. *Emerging Diseases of Animals* 217– 43 (2000).
11.Ridley, A. & Threlfall, E. J. Molecular epidemiology of antibiotic resistance genes in multiresistant epidemic Salmonella typhimurium DT 104. *Microbial drug resistance (Larchmont, N.Y.)* **4**, 113–8 (1998).
12.Chalker, R. B. & Blaser, M. J. A review of human salmonellosis: III. Magnitude of Salmonella infection in the United States. *Reviews of infectious diseases* **10**, 111–24 (1988).
13.Gomez, T. M., Motarjemi, Y., Miyagawa, S., Käferstein, F. K. & Stöhr, K. Foodborne salmonellosis. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales* **50**, 81–9 (1997).
14.Mather, A. E. *et al.* Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. *Science (New York, N.Y.)* **341**, 1514–7 (2013).
15.Revoredo-Giha, C. *et al. Analysis of red and processed meat purchases in Scotland using representative supermarket panel data*. (2009).at <www.foodbase.org.uk//admintools/reportdocuments/338-1-594_S14046.pdf.>
16.Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* **7**, 214 (2007).
17.Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* **29**, 1969–73 (2012).
18.Okoro, C. K. *et al.* Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nature genetics* **44**, 1215–21 (2012).
19.Boyd, D. A., Peters, G. A., Ng, L. & Mulvey, M. R. Partial characterization of a genomic island associated with the multidrug resistance region of Salmonella enterica Typhymurium DT104. *FEMS microbiology letters* **189**, 285–91 (2000).
20.Hall, R. M. Salmonella genomic islands and antibiotic resistance in Salmonella enterica. *Future microbiology* **5**, 1525–38 (2010).
21.Targant, H., Doublet, B., Aarestrup, F. M., Cloeckaert, A. & Madec, J.-Y. IS6100-mediated genetic rearrangement within the complex class 1 integron In104 of the Salmonella genomic island 1. *The Journal of antimicrobial chemotherapy* **65**, 1543–5 (2010).
22.Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M. & Ussery, D. W. Genomic variation in Salmonella enterica core genes for epidemiological typing. *BMC genomics* **13**, 88 (2012).
23.Glynn, M. K. *et al.* Emergence of multidrug-resistant Salmonella enterica serotype typhimurium DT104 infections in the United States. *The New England journal of medicine* **338**, 1333–8 (1998).
24.Threlfall, E. J., Ward, L. R., Frost, J. A. & Willshaw, G. A. Spread of resistance from food animals to man--the UK experience. *Acta veterinaria Scandinavica. Supplementum* **93**, 63–8; discussion 68–74 (2000).
25.Prager, R. *et al.* Clonal relationship of Salmonella enterica serovar typhimurium phage type DT104 in Germany and Austria. *Zentralblatt für Bakteriologie : international journal of medical microbiology* **289**, 399–414 (1999).
26.Threlfall, E. J., Ward, L. R. & Rowe, B. Increasing incidence of resistance to trimethoprim and ciprofloxacin in epidemic Salmonella typhimurium DT104 in England and Wales. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **2**, 81–84 (1997).
27.Low, J. C., Angus, M., Hopkins, G., Munro, D. & Rankin, S. C. Antimicrobial resistance of Salmonella enterica typhimurium DT104 isolates and investigation of strains with transferable apramycin resistance. *Epidemiology and infection* **118**, 97–103 (1997).
28.Ward, L. R., Threlfall, E. J. & Rowe, B. Multiple drug resistance in salmonellae in England and Wales: a comparison between 1981 and 1988. *Journal of clinical pathology* **43**, 563–6 (1990).
29.Witte, W. Medical consequences of antibiotic use in agriculture. *Science* **279**, 996–7 (1998).
30.Rabsch, W., Tschäpe, H. & Bäumler, A. J. Non-typhoidal salmonellosis: emerging problems. *Microbes and infection / Institut Pasteur* **3**, 237–47 (2001).
31.Prager, R., Liesegang, A. & Streckel, W. Salmonellaenterica, serovar Typhimurium, phage type DT104 the emerging epidemic clone in Germany. *Proc 4th Int Meet Bacterial Epidemiological Markers* 104 (1997).
32.Baggesen, D. L. & Aarestrup, F. M. Characterisation of recently emerged multiple antibiotic-resistant Salmonella enterica serovar typhimurium DT104 and other multiresistant phage types from Danish pig herds. *The Veterinary record* **143**, 95–7 (1998).
33.Threlfall, E. J., Rowe, B. & Ward, L. R. A comparison of multiple drug resistance in salmonellas from humans and food animals in England and Wales, 1981 and 1990. *Epidemiology and infection* **111**, 189–97 (1993).
34.Angulo, F. J. & Griffin, P. M. Changes in antimicrobial resistance in Salmonella enterica serovar typhimurium. *Emerging infectious diseases* **6**, 436–8 (2000).
35.Doublet, B., Boyd, D., Mulvey, M. R. & Cloeckaert, A. The Salmonella genomic island 1 is an integrative mobilizable element. *Molecular microbiology* **55**, 1911–24 (2005).
36.Schmieger, H. & Schicklmaier, P. Transduction of multiple drug resistance of Salmonella enterica serovar typhimurium DT104. *FEMS microbiology letters* **170**, 251–6 (1999).
37.Cloeckaert, A. & Schwarz, S. Molecular characterization, spread and evolution of multidrug resistance in Salmonella enterica typhimurium DT104. *Veterinary research* **32**, 301–10 (2001).
38.Cloeckaert, A. *et al.* Occurrence of a Salmonella enterica serovar typhimurium DT104-like antibiotic resistance gene cluster including the floR gene in S. enterica serovar agona. *Antimicrobial agents and chemotherapy* **44**, 1359–61 (2000).
39.Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821–9 (2008).
40.Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy* **67**, 2640–4 (2012).
41.Leekitcharoenphon, P. *et al.* snpTree--a web-server to identify and construct SNP trees from whole genome sequence data. *BMC genomics* **13 Suppl 7**, S6 (2012).
42.Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).
43.Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
44.Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution* **22**, 1185–92 (2005).

National Food Institute
Technical University of Denmark
Mørkhøj Bygade 19
DK - 2860 Søborg

Tel. 35 88 70 00
Fax 35 88 70 01

www.food.dtu.dk

ISBN: 978-87-93109-16-2